VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

ANALÝZA REALITNÍHO TRHU POMOCÍ INFORMACÍ NA INTERNETU
ANALYSIS OF REAL ESTATE MARKET USING INFORMATION ON INTERNET

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Martin Bulín**
AUTHOR

VEDOUCÍ PRÁCE — **doc. Ing. Dan Komosný, Ph.D.**
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Telekomunikační a informační technika**
Ústav telekomunikací

Student: Bc. Martin Bulín ***ID:*** 143828
Ročník: 2 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Analýza realitního trhu pomocí informací na Internetu

POKYNY PRO VYPRACOVÁNÍ:

Naprogramujte aplikaci, která bude prohledávat obsah webových stránek zvoleného realitního portálu v ČR. Ze získaného obsahu separujte informace o realitních kancelářích a o nemovitostech k prodeji. Navrhněte formát pro ukládání získaných informací ve formě textových souborů. Vytvořte rozhraní pro vyhledávání jednotlivých nemovitostí podle zadaných kriterií za účelem jejich srovnávaní pro odhad ceny nemovitostí. Aplikaci sestavte v programovacím jazyce Python.

DOPORUČENÁ LITERATURA:

[1] PILGRIM, M. Ponořme se do Python(u) 3. CZ.NIC, 2010. 435 s. ISBN: 978-80-904248-2-1.

[2] CASTRO. E. HTML, XHTML a CSS: názorný průvodce tvorbou WWW stránek. 1. vyd. Computer Press, 2007. 440 s. ISBN: 978-80-251-1531-2.

Termín zadání: 1.2.2017 ***Termín odevzdání:*** 24.5.2017

Vedoucí práce: doc. Ing. Dan Komosný, Ph.D.
Konzultant:

doc. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Tato práce byla vytvořena jako pomocný nástroj pro odhad ceny nemovitostí porovnávací metodou, pro kterou je potřeba databáze nemovitostí. Tato databáze nemovitostí je vytvářena programem. Diplomová práce má za cíl vytvořit aplikaci v jazyce Python, která bude prohledávat obsah webových stránek zvoleného realitního portálu v ČR. Dalším cílem této práce je vytvořit program pro vyhledávání jednotlivých nemovitostí podle zadaných kriterií, za účelem jejich srovnávání pro odhad ceny nemovitostí. Teoretická část práce je věnována vysvětlení pojmu dolování dat a architektury dolování dat. V praktické části je nejdříve navrhnuta aplikace, která je poté implementována. Vytvořená aplikace prochází webové stránky zvoleného realitního portálu a získává data o realitních kancelářích, jejich inzerátech a fotkách inzerátů. Pro snadnější práci s výstupním souborem, ve kterém jsou uloženy inzeráty, je vytvořena aplikace na vyhledávání jednotlivých inzerátu podle zadané specifikace uživatele. Tento program umožňuje rychlé vyhledávání požadovaných dat ve výstupním souboru. S těmito daty se dále pracuje, je možné je analyzovat a vytvářet zajímavé statistiky a mapy.

KLÍČOVÁ SLOVA

Analýza, Realitní kancelář, Dolování dat, Mozilla Firefox, Webdriver, Selemium, BeautifulSoup, LXML, Python, HTML, Aplikace, Geopy, Mapy API, Skript

ABSTRACT

This thesis was created as a helping tool to estimate the price of properties by comparative method, which needs a database of properties. This database is created by a program. The main aim of this thesis is to create an application in Python language, which will search through contents of websites of a chosen real estate portal in Czech Republic. The next aim of this thesis is to create program which will search for real estate according to chosen criteria. The purpose of these criteria is to compare and estimate the price of properties. In the theoretical part of this thesis I will describe data mining and architecture of data mining. In the practical part I will design an aplication and in the end I will implement it. This aplication will go through websites and obtain data about real estate offices, their advertisements and photos of advertisements. For easier work with the output file that contains the advertisements, I created an application for searching individual advertisements based on users specification. This program allows to quickly search for the requested data from the output file. These data are subject to further work, it is possible to analyse them and create interesting statistics and maps.

KEYWORDS

Analysis, Real estate agency, Data mining, Mozilla Firefox, Webdriver, Selemium, BeautifulSoup, LXML, Python, HTML, Application, Geopy, Mapy API, Script

BULÍN, Martin *Analýza realitního trhu pomocí informací na Internetu*: diplomová práce. BRNO: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2017. 74 s. Vedoucí práce byl doc. Ing. Dan Komosný, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Analýza realitního trhu pomocí informací na Internetu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

BRNO ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu doc. Ing. Danu Komosnému, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

BRNO ....

podpis autora

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této diplomové práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

BRNO ....

podpis autora

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

OBSAH

SEZNAM OBRÁZKŮ

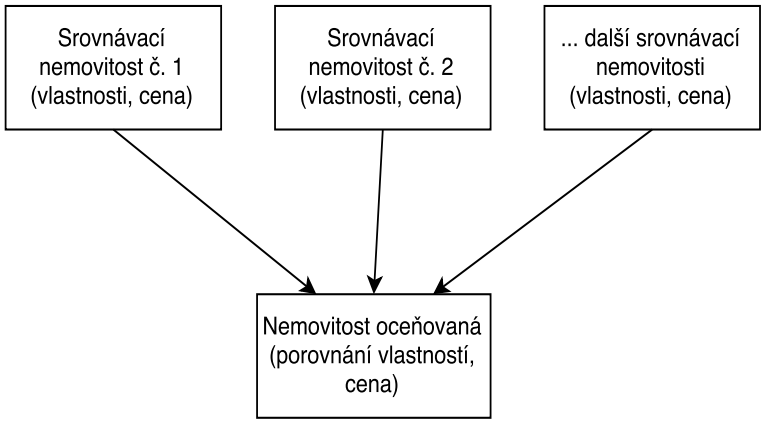
SEZNAM TABULEK

SEZNAM VÝPISŮ

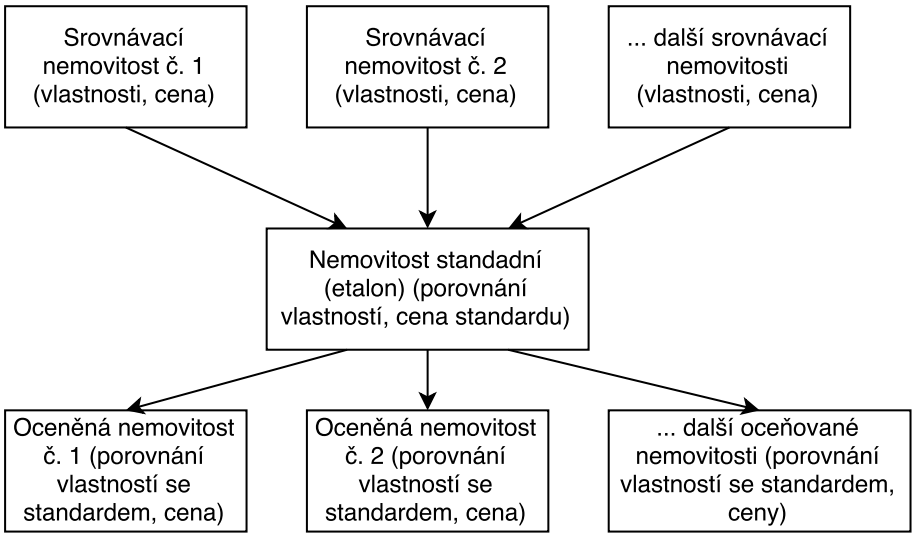
ÚVOD

Tato práce byla vytvořena jako pomocný nástroj pro odhad ceny nemovitostí porovnávací metodou. Diplomová práce má za cíl vytvořit aplikaci v jazyce Python, která bude prohledávat obsah webových stránek zvoleného realitního portálu v ČR. Dalším cílem této práce je vytvořit program pro vyhledávání jednotlivých nemovitostí podle zadaných kriterií. Úkolem je vytvořit databázi srovnávacích nemovitostí za účelem odhadu ceny nemovitostí. Tato databáze bude sloužit k porovnávání nemovitostí. Existují tři hlavní metody používané v realitní praxi, a to metoda porovnávací, výnosová a nákladová. Porovnávací metoda se dále dělí na:

- Metoda přímého porovnávání – U této metody se každá nemovitost z databáze porovná s oceňovanou nemovitostí. Tato metoda je zobrazena na obrázku 1.
- Metoda nepřímého porovnávání – Metoda, při níž je oceňovaná nemovitost porovnávána se standardním objektem podle přesně definovaných vlastností a její cenou. Cena standardního objektu je odvozena na základě zpracované databáze nemovitostí. Tuto metodu lze využít při stanovení ceny více obdobných objektů za využití jedné databáze[1]. Metoda je zobrazena na obrázku 2. Etalon na obrázku je reprezentativní typ oceňování k reprezentativnímu typu. Vybrané srovnávací nemovitosti se použitelnou databází stanou až po utřídění a statickém zpracování dat o vybraných nemovitostech. Součástí databáze by mělo být i uvedení způsobu zjištění dat, aktuálnost a zdroj.



Obr. 1: Metoda přímého porovnání.



Obr. 2: Metoda nepřímého porovnání.

V kapitole 1 je popsána metodologie pro získávání potenciálně užitečných informací z dat nazvaná: technika dolování dat. Dále jsou zde osvětleny oblasti využití této techniky. Kapitola 2 je věnována principu webových stránek, popisu struktury jazyka HTML a HTML značek. V aplikaci jsou použity různé nástroje pro dolování dat. Tyto nástroje jsou rozděleny podle toho, jestli se využívají ve statických webových stránkách, anebo dynamických webových stránkách. Nástroje pro dolování dat ze statických stránek jsou popsány v kapitole 3. Tato kapitola je zaměřena na knihovnu BeautifulSoup, která dokáže získávat informace z HTML dat. Nástroje pro dolování dat z dynamických stránek jsou popsány v kapitole 4. Zde je popsán webdriver Selenium, který dokáže získat HTML data z dynamických webových stránek. Před implementací aplikace jsou v kapitole 5 analyzovány a specifikovány požadavky pro aplikaci. Dále je zde obecný návrh aplikace a návrh objektů pro realitní kancelář a inzeráty. Na závěr je vytvořen vývojový diagram aplikace. Kapitola 6 je věnována popisu implementace aplikace. Jsou zde popsány všechny funkce a principy implementace. Výsledná aplikace dokáže v rámci diplomové práce sbírat data z realitních kanceláří a jejich inzerátů. Dále je v této kapitole popsána implementace programu pro vyhledávání nemovitostí dle zadaných kritérií. V poslední kapitole 7 je provedeno testování obou aplikací a popsáno, s jakými parametry programy pracují. V závěru práce jsou zobrazeny ukázky výstupů obou aplikací a zhodnoceny výsledky práce.

1 DOLOVÁNÍ DAT Z WEBOVÝCH STRÁNEK

1.1 Dolování dat

Pojem data mining neboli dolování v datech bývá definován jako sada automatizovaných postupů, používaných k nalezení dosud neznámých vzorů a vztahů v datech. [2].

Zajímavou definici dolování dat podal v roce 1966 pan Usama Fayad se svými spolupracovníky. Jejich definice říká, že dolování dat je proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty, za účelem získání obchodní výhody [3].

Dolování dat je počítačově softwarová technika získávání dat z webových stránek. Data se mohou z webových stránek získávat strojově nebo ručně. Dolování dat zahrnuje širokou škálu programovacích technik a technologií, jako je analýza dat a informační bezpečnost. V oblastech zabývajících se analýzou a dolováním dat je možné se setkat i s jinými pojmy jako screen scraping, web scraping, web harvesting nebo další. Obecně se však nejčastěji vyskytuje pojem data mining neboli dolování dat, proto tento termín bude využit i v této práci.

Rychlý vývoj technologií v oblasti ukládání dat zaznamenal v posledních letech prudký nárůst. Většina dat se nachází v databázích a jiných datových skladech a mají nestrukturovanou podobu. V těchto datech se nachází velké množství důležitých informací. Dolování v datech je vynikající technika při shromažďování a zpracování velkého množství objemů dat. Úkolem je vyhledávat závislosti na datech a organizovat data do čitelné podoby. Tato technika dokáže zobrazit a prohlížet databáze, které pokrývají tisíce nebo dokonce milióny webových stránek najednou. Pokud by se tato data analyzovala ručně, zabrala by analýza více času i peněz. Trh dolování dat roste v posledních několika letech exponenciální řadou. Ceny technologií se snižují a roste počet instalací. Vznikají vertikální řešení pro jednotlivé obory průmyslu a moderní řízení vztahu se zákazníky je bez techniky dolování dat téměř nemyslitelné [4].

1.2 Oblasti využití techniky dolování dat

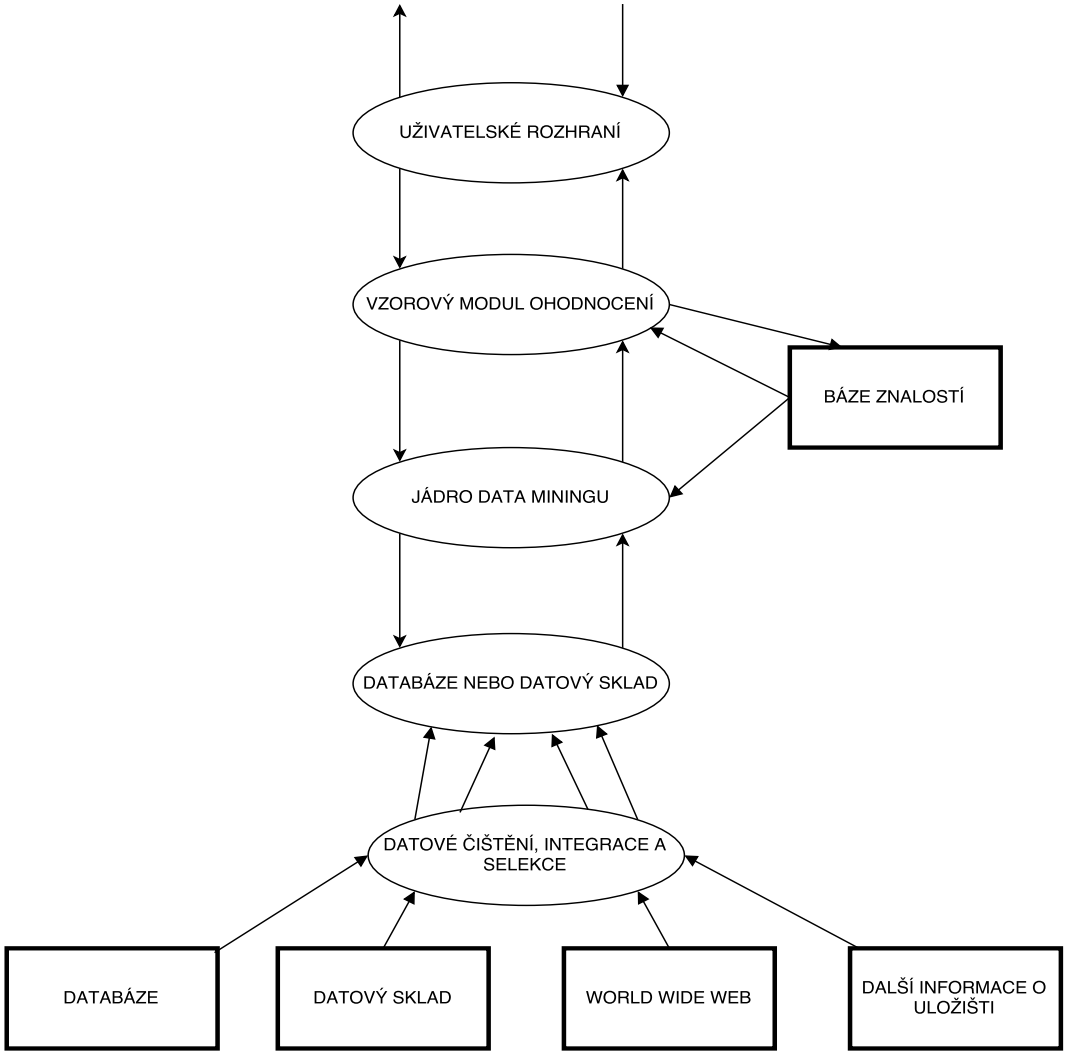
Technika dolování dat v posledních letech přitahuje velikou pozornost nejen v informačním průmyslu, ale i v dalších oblastech průmyslu. Je to logické, jelikož je na internetu k dispozici velké množství dat, která jsou široce dostupná. Z tohoto velkého objemu dat se dají získat zajímavé informace. V současnosti poskytuje technika dolování v datech podniku jedinečnou výhodu před konkurencí a je dnes velice

rozšířenou technologií, která pomůže firmě dosáhnout lepšího postavení na trhu [5]. V dnešní době se technika dolování dat využívá v mnoha oblastech jako například:

- bankovnictví a pojišťovnictví,
- přímý marketing,
- detekce podvodů,
- bioinformatika a lékařství,
- bezpečnost.

1.3 Architektura Dolování dat

Architektura typického systému dolování dat 1.1, může obsahovat následující hlavní komponenty:



Obr. 1.1: Architektura typického systému dolování dat.

- Báze znalostní – Jedná se o znalost webové domény, která slouží k vyhledávání nebo vyhodnocení zajímavých výsledných vzorků dat. Taková znalost může zahrnovat pojem hierarchie, což je organizování atributů nebo hodnot atributů na různých úrovních abstrakce. Další příklady doménových znalostí jsou zajímavé omezení nebo hraniční hodnoty a metadata.
- Technika (jádro) data miningu – Má podstatný význam pro techniku dolování dat. V ideálním případě se skládá ze sady funkčních modulů pro úkoly jako jsou charakterizace, asociace, korelační analýza, klasifikace, předpověď, analýza clusterů (shluků) a analýza vývoje.
- Vzorový modul ohodnocení – Tato složka typicky používá zajímavá opatření a spolupracuje s moduly pro dolování dat tak, aby se soustředily na vyhledávání zajímavých vzorků dat. Tyto vzorky se mohou, za použití hraničních hodnot, filtrovat. Pro efektivní dolování dat je vysoce doporučeno snažit se vyhodnocovat vzorky co nejhlouběji v dolovaném procesu. Díky tomu se hledání omezí na zajímavé vzorky dat.
- Uživatelské rozhraní – Tento modul komunikuje mezi uživateli a systémem pro dolování dat. Uživatel se systémem komunikuje tak, že mu zašle dotaz na dolování dat a poskytuje mu informace, které mohou pomoci vylepšit vyhledávání. Kromě toho, tato komponenta umožňuje uživateli procházet databáze a datová skladiště diagramů nebo datové struktury, vyhodnocovat vydolovaná data a zobrazovat vzorky dat v jiné formě [5].

2 PRINCIP WEBOVÝCH STRÁNEK

Pro získávání dat z webových stránek je potřeba seznámit se s jazykem HTML a HTML značkami. Jazyk HTML je jazykem pro specifikaci rozvržení dokumentů a hypertextových odkazů. Definuje syntaxi a rozmístění speciálních vložených příkazů, které se v prohlížeči přímo nezobrazují, ale které řídí způsob zobrazení obsahu dokumentu, včetně textu, obrázků a ostatních podpůrných médií.

2.1 Struktura dokumentu HTML

Dokumenty HTML se skládají z textu, který definuje obsah dokumentu a ze značek, které popisují jeho strukturu a vzhled. Struktura dokumentu HTML je jednoduchá, tvoří ji vnější značka *<html>*, do níž se dále zapisuje záhlaví a tělo dokumentu. Každý dokument má tedy své záhlaví a tělo, jež po řadě vymezují značky *<head><body>*. Do těla se zapisuje vlastní obsah dokumentu. Tělo zahrnuje zobrazovaný text a označení řídící sekvence (značky) dokumentu, které prohlížeči říkají, jak má daný text zobrazit. Značky se mohou odkazovat také na soubory speciálních efektů, jako jsou grafika a zvuky [6]. Příklad zápisu 2.1 struktury jazyka HTML:

```
<html>
<head>
<title>Ukázka struktury HTML</title>
</head>
<body>
Tento příklad ilustruje <i>jednoduchým způsobem</i>
základní strukturu jazyka HTML
</body>
</html>
```

Výpis 2.1: Příklad zápisu struktury jazyka HTML.

2.2 Značky a atributy

Značky, tedy elementy označené v dokumentech HTML, jsou většinou dobře srozumitelné a snadno se s nimi pracuje, protože je tvoří běžná slova, zkratky a označení. Každá značka se skládá ze jména značky, za níž následuje nepovinný seznam jejich atributů. Jméno značky i její atributy se zapisují mezi úhlové závorky (< a >). Nejjednodušší značku tvoří pouhé jméno značky, zapsané odpovídajícím způsobem

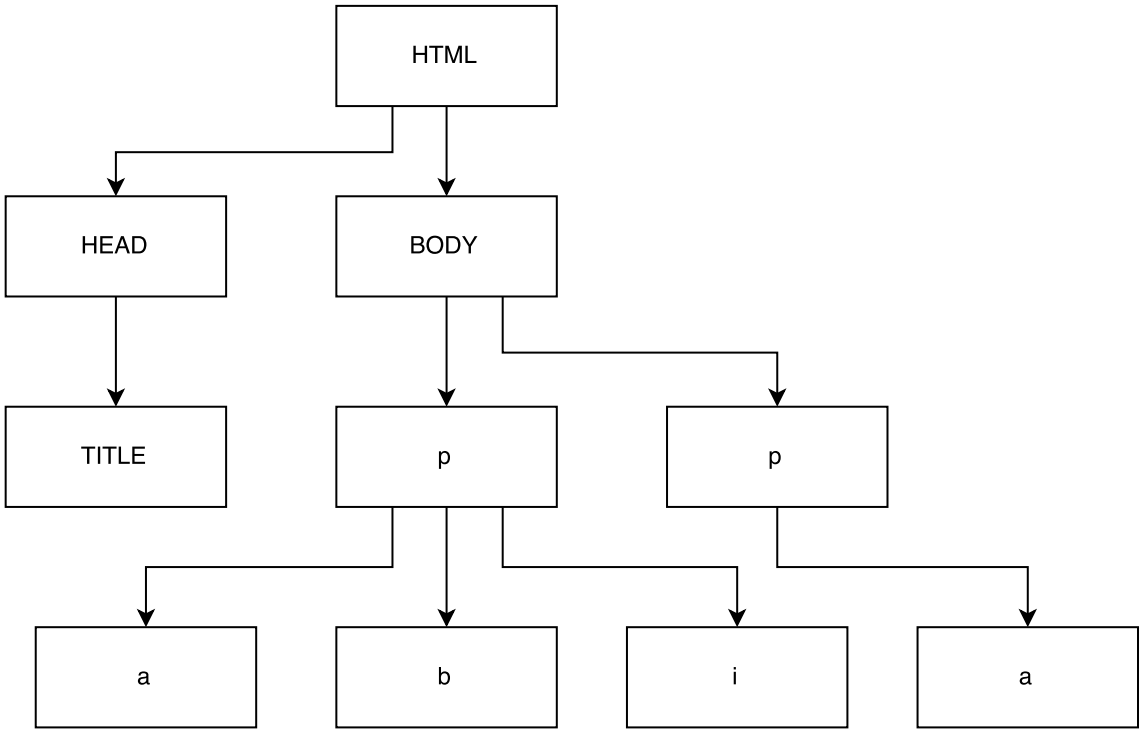
v úhlových závorkách, jako například *<head>*. Složitější značky obsahují jeden nebo více atributů, které blíže popisují nebo modifikují požadované chování značky.

Podle standardu se nerozlišuje velikost písmen ve jménech značek a atributů. Atributy značky spadají za jméno značky a vzájemně se oddělují jedním nebo více znaky tabulátoru. Hodnota atributu značky, je-li definována, se zapisuje za jménem atributu, od něhož se odděluje znaménkem (=). Tvoří-li hodnotu atributu v jazyce HTML jediné slovo, můžeme je zapsat přímo za znaménko (=). Veškeré ostatní hodnoty se musí zapisovat mezi apostrofy nebo uvozovky. Většina prohlížečů toleruje poměrně volný způsob zápisu značek i jejich rozdělení do jednotlivých řádků. Toto pravidlo zvyšuje čitelnost zdrojového textu a snižuje potenciální množství chyb v dokumentu HTML. Níže je ukázka 2.2 výpisu značek v HTML [6]:

```
<a href="http://www.oreilly.com/catalog.tml">
<ul compact>
<ul compact="compact">
<input type=text name=filename size=24 maxlenght=80>
<link title="Obsah dokumentu">
```

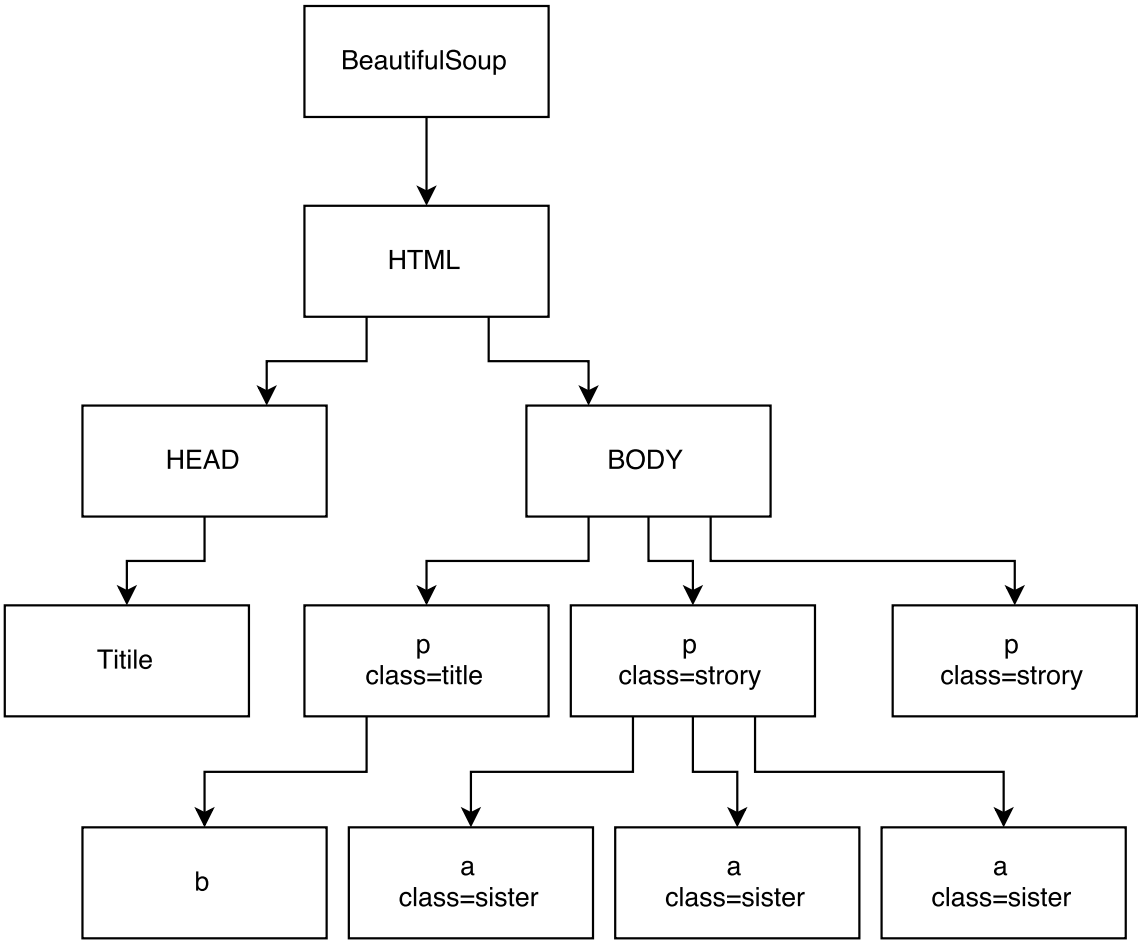
Výpis 2.2: Příklad značek v HTML.

Na obrázku 2.1 je příklad HTML struktury a lze zde vidět, jak mohou být jednotlivé značky seřazeny. Před začátkem každého dolování dat je určitě velice vhodné, zakreslit si danou strukturu do HTML stromu, což zjednoduší získávání informací.



Obr. 2.1: Příklad HTML struktury.

Použití knihovny BeautifulSoup je popsáno na straně 23 a ukázka použití knihovny BeautifulSoup se nachází na výpisu 3.2. Na obrázku 2.2 je vidět k dané ukázce vytvořený HTML strom. Díky tomuto stromu je vyhledávání a dolování informací snadnější. Toto je pouze ukázka. Pokud by se měl nakreslit HTML strom pro některou dynamickou stránku, byl by daleko větší a složitější.



Obr. 2.2: HTML strom k ukázce příkladu 3.2.

2.3 Zobrazení zdrojového kódu webové stránky

Pro pochopení, jak je daná webová stránka strukturovaná, je potřeba zobrazit a prozkoumat její zdrojový kód. Ve většině webových prohlížečů se zobrazí zdrojový kód stránky tak, že se klikne pravým tlačítkem myši na stránku a vybere se volba *Viewed Page Source*. Po tomto příkazu by se měl zobrazit HTML kód stránky. Tento kód se objeví pouze tehdy, jedná-li se o statickou webovou stránku. Pokud by bylo potřeba zobrazit data načtená dynamicky pomocí Javascriptu, musí se označit celá stránka nebo jen elementy, o které je zájem. Poté se klikne pravým tlačítkem myši na výběr a vybere se volba *View Selection Source*.

Další způsob, jak zjistit HTML kód k analýze, je pomocí nástroje Firebug. Označí se elementy, které je potřeba analyzovat a pomocí pravého tlačítka myši se vybere volba *Inspect Element In Firebug*. Nyní je zjištěno, jak získat HTML data k analýze. Tato HTML data se mohou vhodně analyzovat, vybírat si potřebné elementy a

získávat z nich potřebné informace.

2.4 Kaskádové styly

Jednou z klíčových technik používaných v moderním webovém návrhu stránek jsou kaskádové styly (CSS). Je to kolekce metod pro grafickou úpravu webových stránek. Hlavním principem je oddělení prezentace a formátu stránek od obsahu. Dříve byly HTML stránky nepřehledné a obsahovaly několik opakujících se HTML značek. Pokud bylo potřeba změnit jednu značku, musela být změněna ručně.

Proto vznikly kaskádové styly, u kterých je změna značek jednodušší. Kaskádový styl se nachází v odděleném souboru, který doplňuje jednu nebo více HTML stránek na webové stránce. Obsahuje všechny informace o tom jak text, obrázky a rozložení vypadá na webových stránkách[7].

3 DOLOVÁNÍ DAT ZE STATICKÝCH STRÁNEK

Webové stránky se dělí na statické a dynamické. Statická webová stránka zobrazuje pouze statický obsah, proto se používá hlavně v prostředí, kde webová stránka nebude měnit svůj obsah, anebo tam, kde obsah webové stránky může měnit správce webu přímo v kódu HTML. Statické i dynamické webové stránky jsou velice často dokumenty v jazyce HTML, uložené v souborovém systému a dostupné webovým serverem přes protokol HTTP. Statické webové stránky se v dnešní době používají méně často než stránky dynamické.

3.1 Knihovna BeautifulSoup

Knihovna BeautifulSoup je knihovna v jazyce Python, pro získávání dat z jazyka HTML nebo XML. Jazyk Python je moderní, univerzální, skriptovací a programovací jazyk. Knihovna BeautifulSoup je využívána hlavně ve statických webových stránkách. Pro použití knihovny v dynamických webových stránkách je potřeba, nejdříve pomocí různých nástrojů, získat HTML data ukrytá v Javascriptu.

BeautifulSoup rozloží HTML stránku a vytvoří strom, ve kterém je snadné se pohybovat, vyhledávat a modifikovat data. Dále BeautifulSoup pomáhá formátovat, organizovat a opravovat špatně napsaný HTML kód. Jelikož BeautifulSoup není standardní knihovna jazyka Python, musí být před použitím nainstalována. Nejjednodušší je instalace v operačním systému Ubuntu příkazem:

```
$sudo apt-get install python-bs4
```

Nejnovější verze knihovny je BeautifulSoup4. Dále s pomocí systému pip(), který je používán pro instalaci a správu softwarových balíků napsaných v Pythonu, nainstaluji knihovnu BeautifulSoup příkazem:

```
$pip install beautifulsoup4
```

pro python2 a nebo příkazem:

```
$pip3 install beautifulsoup4
```

pro python3. BeautifulSoup podporuje i HTML analyzátory. Jedním z nich je LXML parser, který je v této práci také využíván. Použití knihovny BeautifulSoup je velice snadné a intuitivní. Knihovna se spustí příkazy ve výpisu 3.1

Poté, co se načte knihovna BeautifulSoup v proměnné *html*, bude uložena stránka nebo soubor v jazyce HTML. V posledním kroku se díky LXML parseru získá celý

```
from bs4 import BeautifulSoup
html = BeautifulSoup(open("index.html"))
soup = BeautifulSoup(html, "lxml")
print(soup.prettify())
```

Výpis 3.1: Spuštění knihovny BeautifulSoup.

obsah HTML stránky, který se uloží do proměnné *soup*. Příkazem *print(soup.prettify())* se vytiskne seřazený HTML kód. Pokud se seřazený HTML kód bez problému vytiskne, tak knihovna BeautifulSoup pracuje v pořádku a může se s ní začít pracovat. K práci s knihovnou BeautifulSoup je třeba pochopit problematiku HTML značek (tagů) a jejich atributů. Na výpisu 3.2 je ukázán příklad použití knihovny BeautifulSoup4 [8].

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')

print(soup.prettify())
# <html>
#  <head>
#   <title>
#    The Dormouse's story
#   </title>
#  </head>
#  <body>
#   <p class="title">
#    <b>
#     The Dormouse's story
#    </b>
#   </p>
#   <p class="story">
#    Once upon a time there were three little sisters; and their names were
#    <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#    </a>
#    ,
#    <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#    </a>
#    and
#    <a class="sister" href="http://example.com/tillie" id="link2">
#     Tillie
#    </a>
#    ; and they lived at the bottom of a well.
#   </p>
#   <p class="story">
#    ...
#   </p>
```

```
#   </body>
# </html>
```

Výpis 3.2: Ukázka implementace knihovny BeautifulSoup4.

3.2 Knihovna LXML

Pro Python existuje spousta knihoven pro zpracování HTML a XML řetězců. V této práci je používána knihovna LXML. LXML je open source knihovna, která je vybudována nad populárním parserem libxml2. LXML je knihovna pro Python a obsahuje velké množství funkcí na zpracování HTML a XML. Na rozdíl od ostatních knihoven je rychlejší, podporuje i komplikovanější výrazy a má dobře popsanou dokumentaci. Instalace se provede příkazem na výpisu 3.3:

```
$ apt-get install python-lxml
$ easy_install lxml
$ pip install lxml
```

Výpis 3.3: Instalace knihovny LXML.

Níže na výpisu 3.4 je vypsaná ukázka využití knihovny LXML [9].

```
from lxml import etree
tree = etree.parse('examples/feed.xml')
root = tree.getroot()
root.findall('{http://www.w3.org/2005/Atom}entry')
```

Výpis 3.4: Ukázka využití knihovny LXML.

4 DOLOVÁNÍ DAT Z DYNAMICKÝCH STRÁNEK

V dnešní době se většinou setkáváme s dynamickými webovými stránkami než se stránkami statickými. Dynamická webová stránka generuje obsah webové stránky aktuálními informacemi pro každé individuální zobrazení. Obsah dynamické webové stránky se mění v závislosti na čase, uživateli, kontextu nebo uživatelské interakci. Dynamické webové stránky velice často používají skriptovací jazyky, které jsou provozovány v samotném prohlížeči.

Ve většině případů existují pouze dva jazyky: ActionScript a Javascript. ActionScript je využíván k přehrávání flash aplikací na webu, streamování multimediálních souborů, hraní online her a další. V ActionScriptu se většinou dolování dat nevyužívá, proto se práce zaměří na jazyk, který se využívá v dynamických webových stránkách a tím je Javascript. Jazyk Javascript je popsán v sekci 4.1.

4.1 Skriptovací jazyk Javascript

S dalším programovacím jazykem, se kterým se v projektu pracuje, je Javascript. Javascript je programovací jazyk, který je součástí webových prohlížečů a má za úkol generovat dynamický obsah webových stránek. JavaScript lze použít pouze v internetovém prohlížeči, ovšem na různých místech. Javascript je obsažený přímo v HTML stránce. Protože je označen značkou <SCRIPT>, text skriptu se nezobrazí uživateli, ale je zpracován prohlížečem, který provede požadovanou akci. Značka <SCRIPT> bývá umístěna v hlavičce dokumentu, ale stejně tak je možné ji umístit do těla stránky. Dokument může obsahovat několik párů značek <SCRIPT>. Důležité je, že všechny části kódu roztroušené po dokumentu dávají dohromady jeden program [10].

Pro dolování dat z moderních webových stránek, se musí použít alespoň jedna z knihoven, které se specializují na zobrazení stránek v jazyce Javascript. Mezi takové knihovny patří například jQuery, Selenium nebo PhantomJS.

Při dolování dat z dynamických stránek se stává, že obsah HTML kódu z prohlíženého webu, nesouhlasí se zdrojovým kódem na tomto webu. Jedním z řešení je dolovat data přímo z Javascriptu. Druhým řešením je použít balíčky v jazyce Python, schopné Javascript vykonávat a dolovat data ze stránky tak, jak je vidí webový prohlížeč [11].

4.2 Knihovna Selenium

Selenium je nástroj (API) pro dolování dat, který byl původně vytvořen pro testování nebo automatizaci webových stránek. V dnešní době bývá také často využíván pro přesné zobrazení webových stránek tak, jak se objevují v prohlížeči. Tímto nástrojem je možno získat data generované v Javascriptu. Selenium pracuje automaticky pomocí webového prohlížeče, díky němuž načítá webové stránky. Ze stránek se získávají požadovaná data a to i tak, že je možné sledovat živě, které stránky Selenium přesně prohlíží, nebo které akce zrovna provádí, v okně prohlížeče. Další užitečnou funkcí Selenia je kliknutí na libovolný odkaz, anebo možnost dokázat se vrátit na předchozí stránku.

Selenium neobsahuje vlastní webový prohlížeč. Spuštění Selenia vyžaduje integraci s jinými webovými prohlížeči například Mozilla Firefox, Safari, Opera, Chrome atd.. Aplikace bude pracovat s webovým prohlížečem Mozilla Firefox nebo zkráceně Firefox. Při zpracování se vyskytl problém. Selenium nepodporuje některé verze Firefoxu. Jedná se o Firefox verze 47.0. Vytvořená aplikace pracuje s Mozillou Firefox verze 45.0, pod kterou Selenium funguje. Jelikož je spousta verzí internetového prohlížeče, není zaručeno, že tento program bude správně fungovat na všech verzích. Pro spuštění Selenia se správnou verzí Firefoxu byl vytvořen skript s názvem *local-Firefox.py*, který bude popsán v kapitole 6. Selenium se skládá ze čtyř komponent:

- Selenium IDE – Je integrované vývojové prostředí pro Selenium skripty. Je implementován jako rozšíření Firefoxu a umožňuje nahrávat, upravovat a ladit testy.
- Selenium client API – Alternativa psaní testů v Selenese, testy mohou být napsané v jiných programovacích jazycích. Tyto testy pak komunikují se Seleniem pomocí volání metod v Selenium client API.
- Selenium Grid – Selen Grid je server, který umožňuje testům použít instance webového prohlížeče běžící na vzdálených počítačích. S vydáním Selenia 2.0 má Selenium Server již zabudovanou Grid funkcionalitu.
- Selenium Remote Control – Je testovací nástroj, který umožňuje psát automatizované UI testy webových aplikací v libovolném programovacím jazyce.
- Selenium Webdriver – Hlavní změnou v Seleniu Webdriveru je integrace WebDriver API. Webdriver je navržen pro co nejjednodušší a nejstručnější programovací rozhraní. Selenium–Webdriver byl vyvinut pro lepší podporu dynamických webových stránek, kde se mohou elementy měnit bez nutnosti znovu načtení celé stránky. Webdriver si klade za cíl poskytnout dobře navržené a objektově orientované API, které poskytuje lepší podporu pro moderní problémy testování webových aplikací [12]. Webdriver podporuje operační systémy Microsoft Windows, Apple OS X a Linux. Programovací jazyky jsou

podporovány prostřednictvím Selenium ovladačů. Jedná se o knihovny vytvořené pro každý jazyk, které převádí příkazy ze Selenium API do formy metod a funkcí. Selenium podporuje spoustu programovacích jazyků například Java, Python a další.
Pokud bylo potřeba preferovat běh skriptů na pozadí bez spuštěného prohlížeče, je možné použít nástroj PhantomJS. PhantomJS je nástroj, který načítá webové stránky v paměti. Pokud by vadil otevřený prohlížeč, je PhantomJS velmi užitečný nástroj. Selenium–Webdriver se jeví jako vhodný nástroj pro analýzu a dolování dat z webových stránek. Zobrazí celou stránku v HTML kódu i to, co je obsaženo v Javascripu. Dále Selenium–Webdriver podporuje nástroje, se kterými bude aplikace pracovat. Těmito nástroji jsou operační systém Linux a programovací jazyk Python, který má knihovnu BeautifulSoup. Knihovna BeautifulSoup je popsána v kapitole 3.1.

5 ANALÝZA, SPECIFIKACE POŽADAVKŮ A NÁVRH APLIKACE

Diplomová práce má dva nosné cíle. Prvním cílem je vytvořit aplikaci *diplomka.py*, která bude prohledávat obsah webových stránek zvoleného realitního portálu v ČR. Ze získaného obsahu se budou získávat a separovat informace o realitních kancelářích a o nemovitostech k prodeji. Výstupem aplikace bude složka pojmenovaná *output_time*, kde time je časová značka začátku programu. Tato složka bude obsahovat dva textové soubory s informacemi o realitních kancelářích a jejich inzerátech. Dále bude složka obsahovat i soubory s uloženým zdrojovým kódem stránky a fotkami inzerátů.

Druhý cíl je realizován aplikací *search.py*. Úkolem této aplikace je vytvořit program pro vyhledávání jednotlivých nemovitostí. Nemovitosti se vyhledávají podle zadaných kritérií uživatele. Účelem vyhledávání nemovitostí je jejich srovnání pro odhad ceny nemovitostí porovnávací metodou. Obě aplikace jsou vytvořeny v programovacím jazyce Python pro operační systém Linux. V předchozí kapitole jsou popsány nástroje pro dolování dat, které jsou potřeba pro sestavení programu. V této kapitole je aplikace podrobně specifikovaná a navrhnuta.

5.1 Specifikace zadání aplikace

Aplikace *diplomka.py* by měla být schopna pracovat nad zvoleným realitním portálem.

- Prvním úkolem aplikace *diplomka.py* bude získat data o realitních kancelářích. Každá realitní kancelář bude identifikována jednoznačným identifikátorem a měl by k ní být přiřazen čas jejího uložení. Čas je vhodné pro další práci uložit do formátu ISO. Dalším bodem zadání bude zjistit o každé realitní kanceláři další informace, jako jsou název realitní kanceláře, adresa, telefon, email, webová adresa, počet inzerátů, počet makléřů, počet poboček, web na serveru Sreality, stát, kraj, nejpočetnější město, zeměpisné souřadnice nejpočetnějšího města, zeměpisné souřadnice realitní kanceláře, počet inzerátů v nejčastějším městě a počet všech měst, ve kterých se nemovitosti nacházejí. Položky realitní kanceláře budou od sebe odděleny tabulátorem a uloží se do souboru s názvem *realEstateAgency_time.txt*, kde proměnná time je čas začátku programu.
- Druhým úkolem aplikace *diplomka.py* bude zjistit informace o jednotlivých inzerátech nemovitostí, které realitní kanceláře nabízejí. Každý inzerát by měl obsahovat i údaje od své realitní kanceláře, která ji nabízí. Těmito údaji jsou ID realitní kanceláře, čas, kdy byla realitní kancelář uložená a název realitní

kanceláře. Stejně jako u realitních kanceláří bude i inzerátu přiřazen jedinečný identifikátor a čas uložení nemovitosti. Čas u nemovitosti bude také ve formátu ISO. Další povinné údaje, které podle specifikace zadání bude inzerát obsahovat, jsou název inzerátu, adresa, cena, jméno a příjmení makléře, který inzerát nabízí, jeho adresa, webové stránky, mobil, email, text inzerátu a na konci by se měla uložit zjištěná poloha inzerátu. Dalším bodem zadání bude zjistit nepovinné položky nemovitosti. Každý inzerát nemovitosti může obsahovat libovolný počet položek, jako je například celková cena, ID zakázky nebo energetická náročnost budovy. Všechny údaje budou odděleny tabulátorem a uloží se do souboru *inzeraty_time.txt*, kde proměnná time je čas začátku programu.

Pokud by se u dané položky nevyskytoval její údaj, vloží se tam místo určité hodnoty slovo *"NA"*. Naopak, pokud by měla položka údajů více, vloží se do tabulky pouze údaj, který položka obsahovala jako první. Na začátku každého souboru musí být hlavička, která bude obsahovat všechny hledané položky. Pro lepší zobrazování, například v Excelu, se před každou položku vloží znak *"#"*. Oba dva soubory budou ve formátu *"txt"*. Při spuštění by se měla na ploše počítače vytvořit složka s názvem, který si uživatel zvolí v parametru. Tato složka se vytvoří v místě spuštění programu. Do této složky se vloží dvojice souborů.

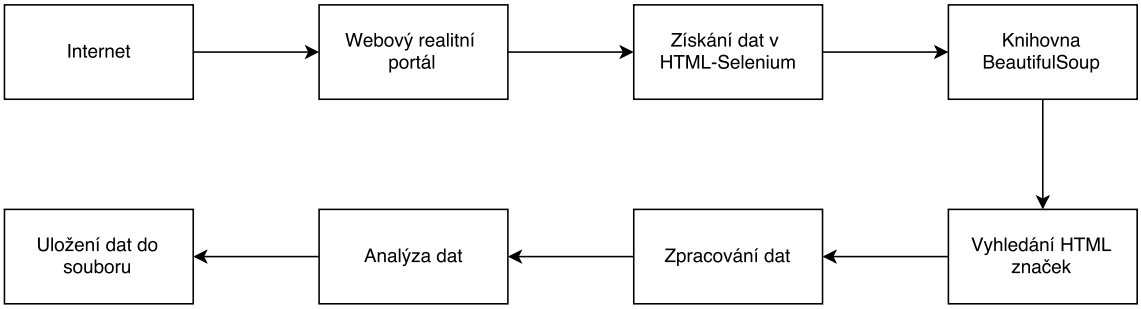
Ve vytvořené složce by se dále měly nacházet adresáře pro každou realitní kancelář. Tyto adresáře by měly být pojmenované podle jednoznačného identifikátoru realitní kanceláře. Každý jednotlivý adresář bude obsahovat soubor se svojí HTML stránkou. Pokud daná realitní kancelář obsahuje inzeráty, bude obsahovat i adresář pro každý inzerát. Složky budou pojmenovány podle jednoznačného identifikátoru inzerátu realitní kanceláře a měly by obsahovat soubor, který obsahuje kód HTML stránky daného inzerátu a jeho obrázky.

Posledním úkolem bude vytvořit aplikaci *search.py*, která bude se staženou databází pracovat. Tato aplikace bude schopna ve výstupních souborech vyhledávat inzeráty podle specifikací zadaných uživatelem. Hlavními kritérii, podle kterých by se mělo vyhledávat, bude vyhledávání podle času, ceny, lokality, typu nemovitosti, vzdálenosti a klíčového slova. Účelem je odhad ceny nemovitostí porovnávací metodou. Další výhodou této aplikace bude rychlejší vyhledávání ve stažené databázi nemovitostí.

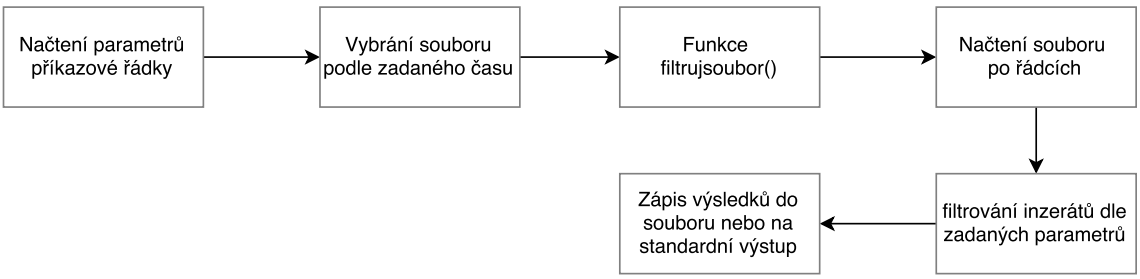
Specifikace aplikace je velice rozsáhlá, proto je důležité, vytvořit si nejdříve návrh aplikace.

5.2 Obecný návrh aplikace

Před implementací aplikace je výhodné si aplikaci vhodně navrhnout. Na obrázku 5.1 je vytvořen obecný návrh aplikace pro získávání dat. Je zřejmé, že aplikace bude muset zvládnout několik úloh. Nejprve se bude dotazovat na stránku zvoleného realitního portálu. Pomocí nástroje Selenium se přesune na zvolenou stránku, ze které získá HTML data. Tato data se zpracují pomocí knihovny BeautifulSoup, pomocí jednotlivých HTML značek. Získaná data se analyzují a vyberou se informace, které jsou zapotřebí pro danou aplikaci. Nakonec se získané informace vloží do výstupní struktury a uloží se do souboru. Na obrázku 5.2 je obecný návrh aplikace pro vyhledávání získaných dat.



Obr. 5.1: Obecný návrh aplikace pro získávání dat.

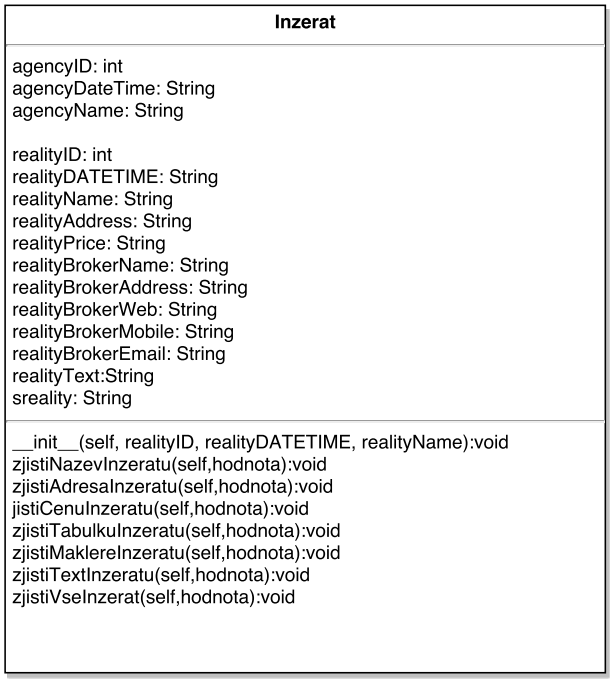


Obr. 5.2: Obecný návrh aplikace pro vyhledávání získaných dat.

5.3 Objekty aplikace

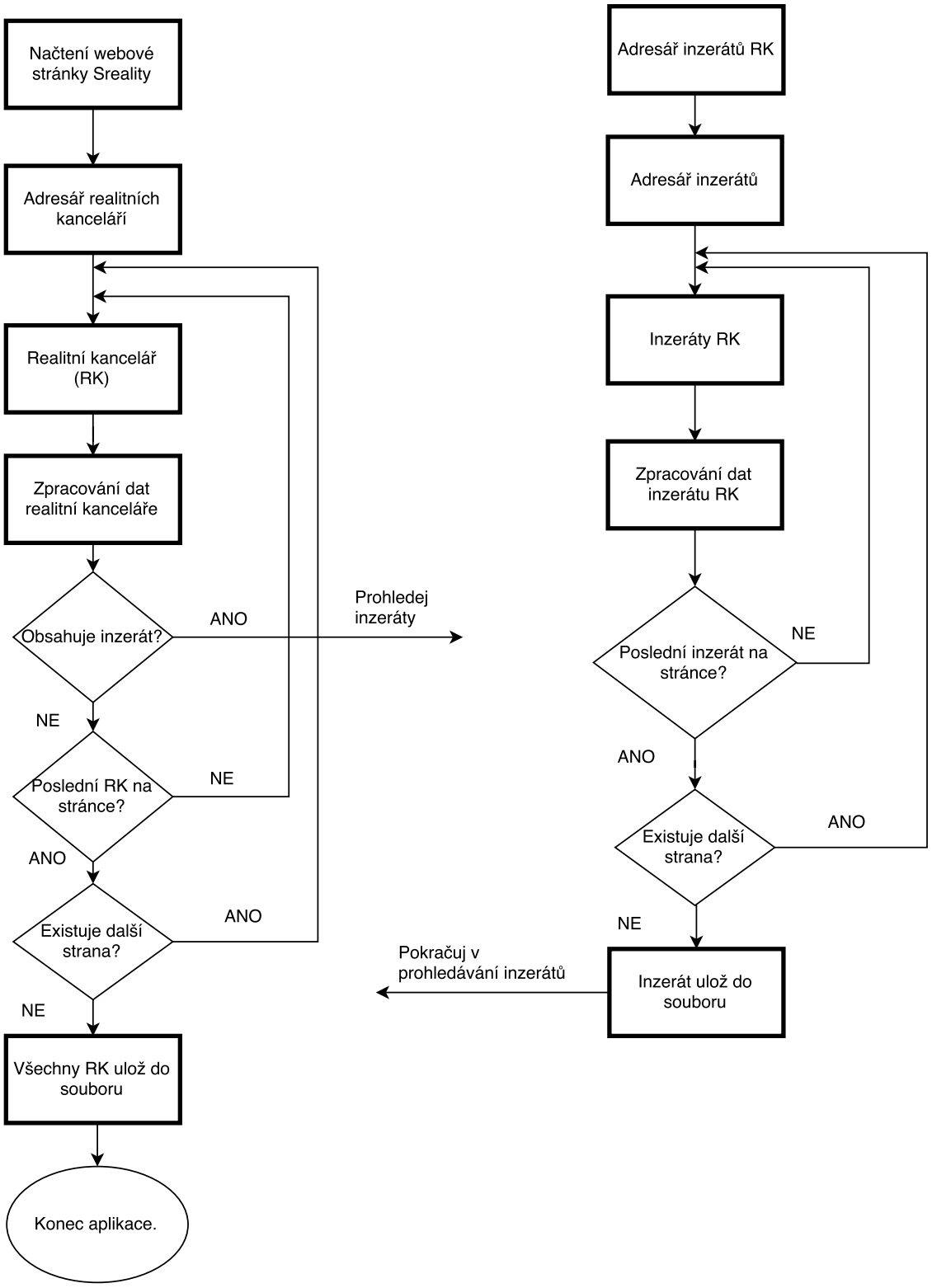
Aplikace by měla zpracovávat informace o realitních kancelářích a inzerátech realitních kanceláří. V programu je využito objektově orientované programování. Vytvoří se dva objekty, objekt realitní kanceláře a objekt pro inzerát. Jakou mají vytvořené objekty strukturu, lze vidět na obrázcích 5.3 a 5.4. Objekty jsou užitečné pro přehlednost, efektivitu a možné budoucí úpravy programu.

Obr. 5.3: Objekt pro realitní kancelář.



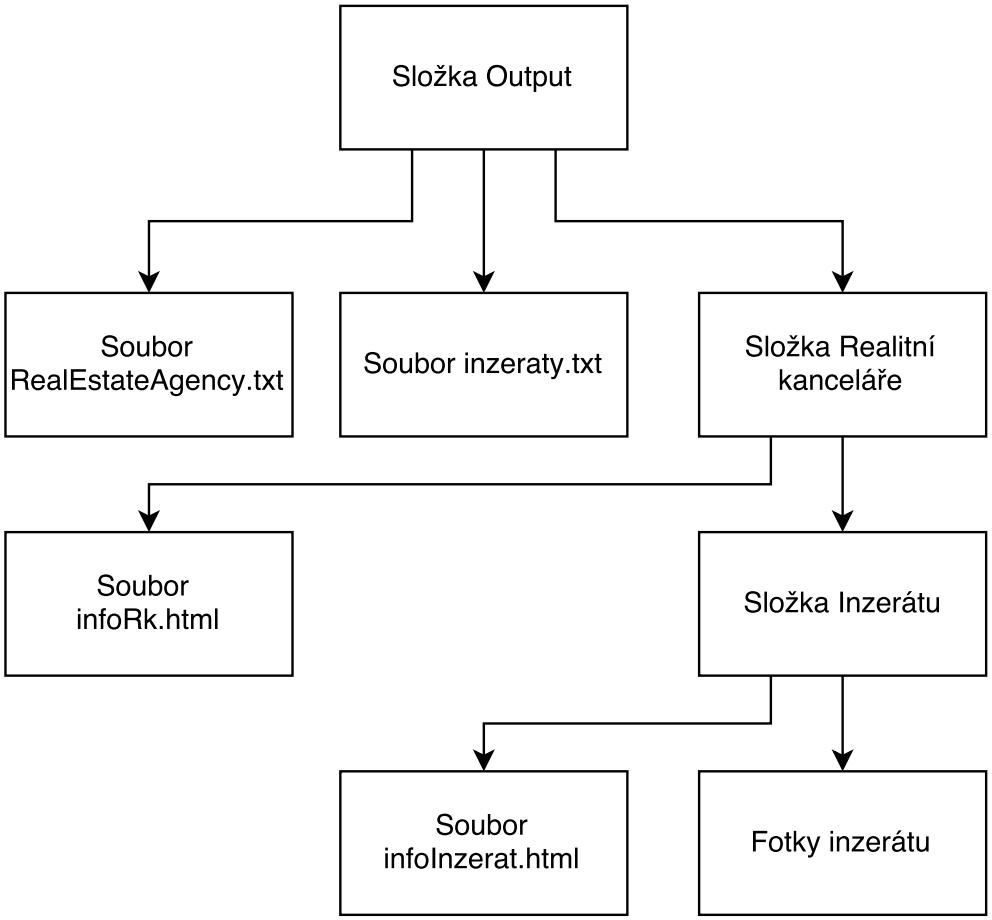
Obr. 5.4: Objekt pro inzerát realitní kanceláře.

5.4 Vývojový diagram aplikace



Obr. 5.5: Vývojový diagram aplikace.

Dále je zmíněno, jak bude vypadat adresářová struktura výsledného programu. Na obrázku 5.6 je vidět, v jaké podobě budou složky, fotky a soubory uloženy ve výstupním programu.



Obr. 5.6: Ukázka adresářové struktury výsledného souboru.

6 IMPLEMENTACE APLIKACÍ

Tato kapitola je zaměřena na implementaci aplikace a je implementovaná pro realitní portál Sreality.cz. Pokud by bylo potřeba využít jiný realitní portál, implementace by byla podobná. Podobnost by byla v implementovaných funkcích aplikace, hledaných informacích a využití Selenia. Rozdíl by byl v HTML značkách, kdy každá webová stránka používá jiné HTML značky. Popis implementace je rozdělen na tři části. První část se zabývá implementací zpracování realitních kanceláří. Druhá část je věnována implementaci zpracování inzerátů realitních kanceláří. Ve třetí části bude vytvořen program, který z výsledných sesbíraných dat, bude vyhledávat uživatelem definované inzeráty podle zadaných parametrů. Vstupem aplikace budou nestrukturovaná HTML data. Výstupem aplikace budou nalezená strukturovaná data v textových souborech.

6.1 Získávání a separování dat z webových stránek realitních kanceláří

Před začátkem implementace aplikace se nejdříve vloží všechny knihovny, které bude aplikace využívat. Jsou to například BeautifulSoup, Selenium, Geopy, knihovny pro určení identifikátoru, datumu, času, vytvoření složek atd. Dále se načte webdriver Selenium tak, aby fungoval na internetovém vyhledávači Firefox. Jelikož je spousta verzí internetového prohlížeče, nelze zaručit, že tento program bude správně fungovat na všech verzích. Proto je vytvořen skript s názvem *localFirefox.py*. Tento skript stáhne určitou verzi vyhledávače Firefox a rozbalí ji v místě spuštění programu. Poté lze hlavní program spustit ze složky, která obsahuje stažený Firefox, pomocí webdriveru Selenium příkazem:

```
binary = FirefoxBinary(os.path.expanduser('~')+'/Firefox/firefox/
    firefox-bin')
driver = webdriver.Firefox(firefox_binary=binary)
driver.maximize_window()
driver.get("https://www.sreality.cz/adresar")
```

Výpis 6.1: Spuštění webdriveru Selenium.

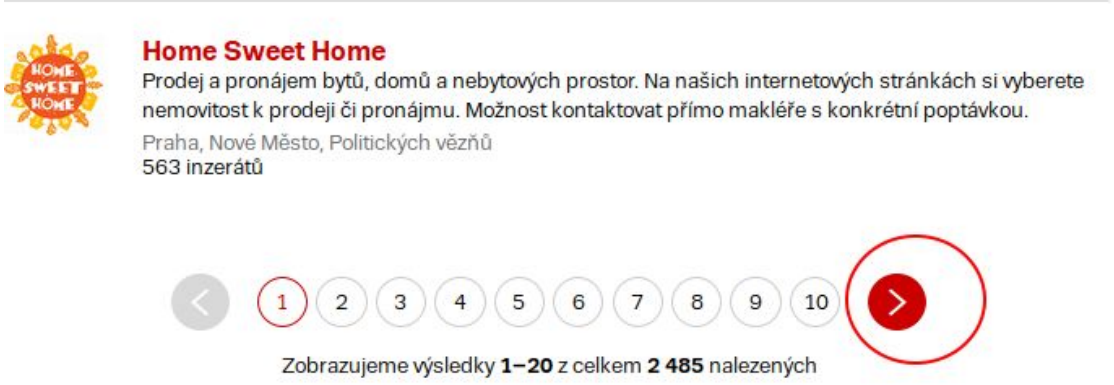
Nyní se v hlavní funkci programu nejdříve zpracují argumenty příkazového řádku, se kterými je možné program spouštět. Program lze spouštět s následujícími parametry příkazového řádku:

- *-h* nebo *--help* – Tento parametr vypíše nápovědu k programu na standardní výstup. Může být zadán pouze jednou a nesmí se kombinovat s žádným jiným parametrem.
- *--output* – Tento parametr definuje název výstupního adresáře, do kterého se budou vkládat všechna data. Pokud adresář neexistuje, vytvoří se v místě spuštění programu. Tento parametr může být zadán pouze jednou a může být kombinován s parametry *--realEstateAgency*, *--realEstate* a dalšími.
- *--realEstateAgency* – Tento parametr definuje, kolik má program zpracovat realitních kanceláří v adresáři na portálu Sreality. Pokud není tento parametr zadán, zpracovávají se všechny realitní kanceláře. Tento parametr je vytvořen pro testování programu.
- *--realEstate* – Tento parametr definuje, kolik má program zpracovat inzerátů pro každou realitní kancelář. Pokud není tento parametr zadán, zpracovávají se všechny inzeráty realitních kanceláří. Tento parametr je vytvořen pro testování programu. Pokud je spuštěn program s parametry *--realEstateAgency=1 --realEstate=1*, prohledává se pouze první realitní kancelář a jeden její inzerát.
- *--photo* – Pokud je zadán tento parametr, jsou získány z každého inzerátu i fotky, které obsahuje. Tyto fotky se uloží do složky vytvořené pro každý inzerát.
- *--singleRealEstateAgencyByName* – Pokud je zadán tento parametr, program provede vyhledání jedné zadané realitní kanceláře podle názvu. Tento parametr může být zadán pouze jednou. V případě, kdy je potřeba vyhledat více realitních kanceláří, použije se parametr *--setRealEstateAgencyByName*.
- *--setRealEstateAgencyByName* – Pokud je zadán tento parametr, program provede vyhledání více zadaných realitních kanceláří podle jejich názvu.
- *--singleRealEstateAgencyByUrl* – Pokud je zadán tento parametr, program provede vyhledání jedné zadané realitní kanceláře podle její webové adresy na serveru Sreality. Tento parametr může být zadán pouze jednou.
- *--setRealEstateAgencyByUrl* – Tento parametr má za úkol načíst soubor ve formátu CSV. V tomto souboru se nachází seznam www serverů realitních kanceláří. Z tohoto souboru se vyberou webové adresy realitních kanceláří, které nejsou hostovány a zároveň nenabízejí více než 400 inzerátů. Pro tyto realitní kanceláře se provede sběr inzerátů.

6.1.1 Zpracování informací o realitních kanceláří bez parametru

Pokud byl program spuštěn bez parametrů, budou se procházet všechny realitní kanceláře. Pomocí metody *driver.get* se program sám naviguje na jakékoliv webové

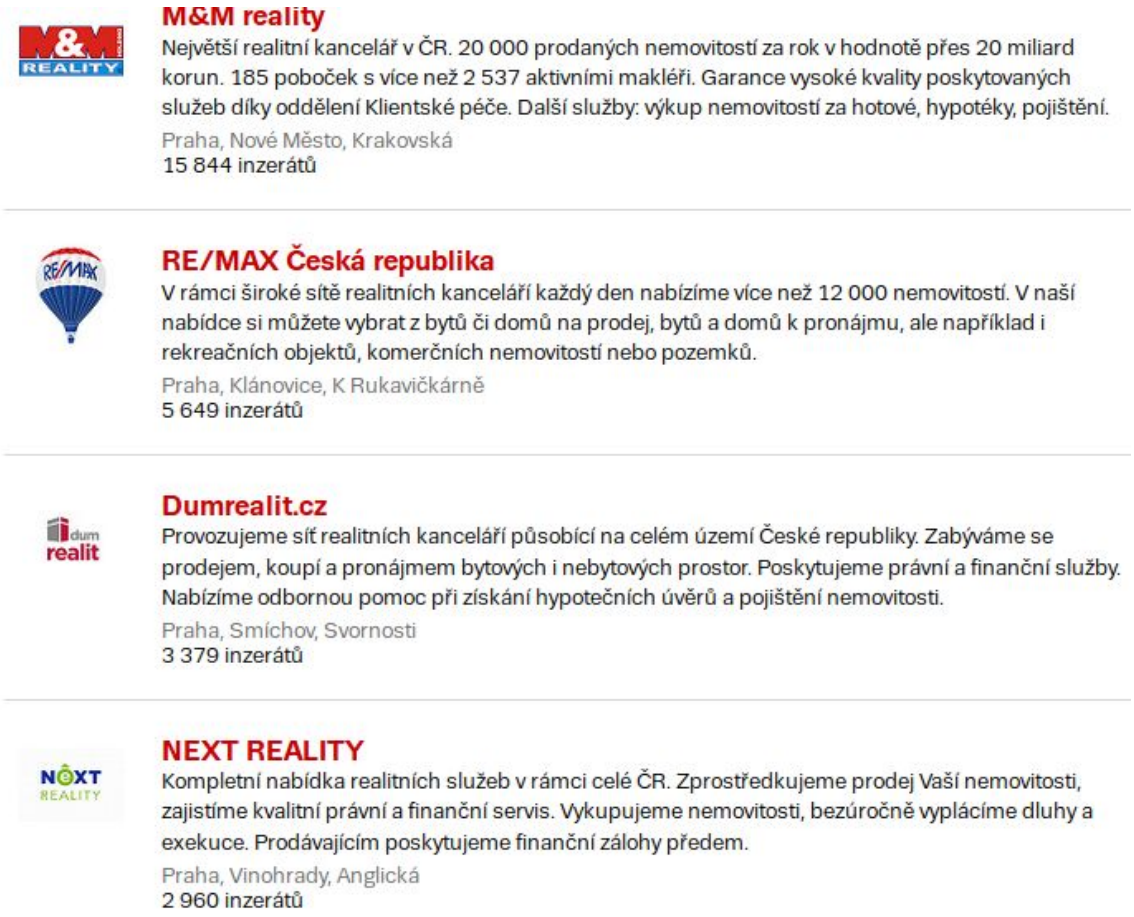
stránky. Příkazem *driver.get("https://www.sreality.cz/adresar")* se otevře stránka, která obsahuje adresář a informace o všech realitních kancelářích. Poté je vytvořen jeden cyklus *while*, který kontroluje, zda je na stránce seznamu realitních kanceláří odkaz na další stránku tak, jak je to uvedené na obrázku 6.1. Pokud se na stránce odkaz na další stránku nachází, přeskočí program na další stránku. Pokud se zde odkaz nenachází, aplikace zjistí, že je to poslední strana. Na této straně program projde poslední realitní kanceláře, ukončí se program a do souboru se zapíší zjištěné výsledky. Pokud se u Selenia vyskytne nějaký problém, je vytvořen *while* cyklus, který má 5 průchodů. Tento cyklus se pokusí, při neúspěšném načtení stránky, stránku znovu načíst. Čekání Selenia na načtení webové stránky je implicitně nastaveno na 30 sekund. Pokud se ani poté nepodaří získat elementy na webové stránce, přechází se na zpracování další realitní kanceláře.



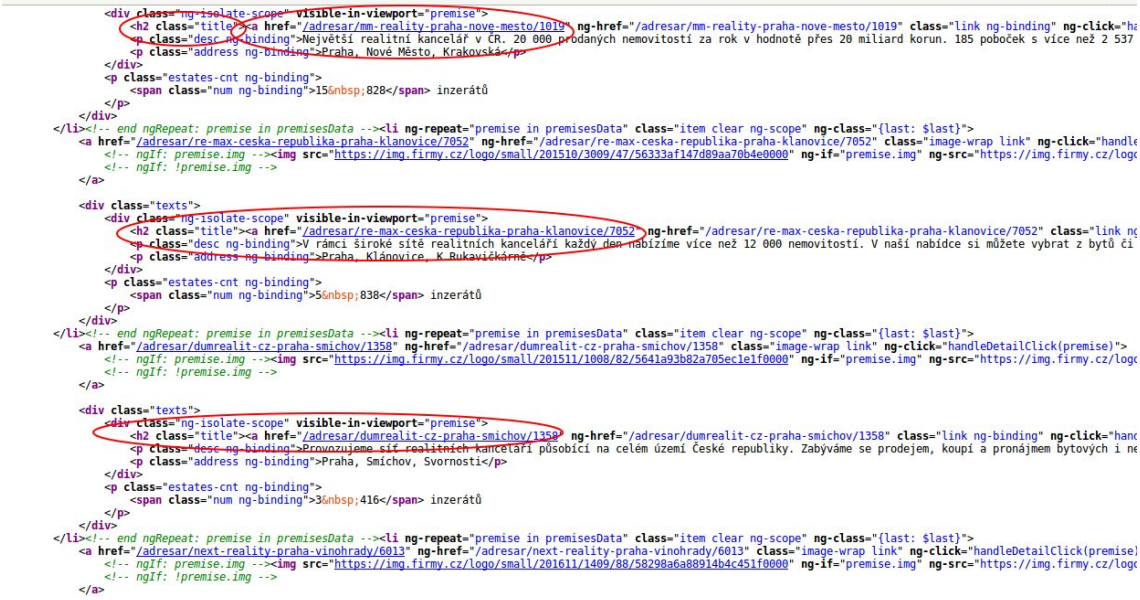
Obr. 6.1: Navigace na další stranu seznamu realitních kanceláří.

V hlavním cyklu *while* se nachází dva cykly *for*. První cyklus *for* vytvoří objekt pro každou realitní kancelář na dané stránce. Do tohoto objektu je vložen jeho jednoznačný identifikátor, čas uložení a jméno realitní kanceláře. Pro každý objekt realitní kanceláře je vytvořena složka s jejím ID. Na obrázku 6.2 je možno vidět ukázku seznamu realitních kanceláří. Program projde každou realitní kancelář v seznamu. K tomu je využita značka v HTML kódu stránky. Na detail realitní kanceláře se program dostane pomocí značek *<h2>* a *<a href>*. Pomocí těchto značek jsou zjištěny všechny webové adresy na detailní stránky realitních kanceláří. Na obrázku 6.3 je vidět, jak jsou tyto značky umístěny a zobrazeny v HTML kódu. Webové adresy jsou uloženy do vytvořeného seznamu s názvem *seznamAdresRk*. Tento seznam umožňuje navigaci pomocí Selenia na všechny detaily stránek realitních kanceláří. Tento seznam se využije i ve druhém cyklu *for*. V minulosti se stávalo, že než program stihl zpracovat všechny inzeráty dané realitní kanceláře, přesunula se některá

realitní kancelář na další stránku a začala se zpracovávat znovu. Proto bylo vytvořeno pole *zpracovaneRk*, do kterého jsou ukládány všechny adresy a zkontrolovány, zda již nebyly jednou zpracovávány.



Obr. 6.2: Ukázka seznamu realitních kanceláří.



Obr. 6.3: HTML značky na stránce Sreality.cz, potřebné pro navigaci v Seleniu.

Druhý cyklus *for* prochází seznam uložených webových stránek realitních kanceláří. Pomocí Selenia se program dokáže dostat na danou stránku uloženou v seznamu. Zde se nachází všechny informace o dané realitní kanceláři tak, jak je zobrazeno v ukázce na obrázku 6.4. Na tomto obrázku jsou poznačené informace, které je potřeba získat. V tomto cyklu je získána celá webová stránka ve formátu HTML. Stránka je vložená do proměnné *html* a uložena do složky s ID dané realitní kanceláře. Nyní je načtena a použita knihovna BeautifulSoup a také knihovna LXML. Získaná data jsou uložena do proměnné *soup*. Příkaz poté vypadá následovně *soup = BeautifulSoup(html,"lxml")*.

Obr. 6.4: Otevřený odkaz na realitní kancelář s vyznačenými informacemi.

V další části jsou načteny funkce *zjistiVseRk()* a *informace()*. Ve funkci *zjistiVseRk()* jsou definované všechny pomocné funkce. Pomocné funkce definované ve funkci *zjistiVseRk()* jsou:

- *zjistiNazev()* – Funkce pro nalezení názvu realitní kanceláře.
- *zjistiAdresa()* – Funkce pro nalezení adresy realitní kanceláře.
- *zjistiWeb()* – Funkce pro nalezení webové adresy realitní kanceláře.
- *zjisiEmail()* – Funkce pro nalezení emailové adresy realitní kanceláře.
- *zjistiTelefon()* – Funkce pro nalezení telefonního čísla realitní kanceláře.
- *getAgencyLongLant()* – Funkce pro nalezení zeměpisné délky a šířky realitní kanceláře z adresy realitní kanceláře.
- *getNejPocetnejsi()* – Funkce pro nalezení zeměpisné délky, šířky a kraje nejčastějšího města, ve kterém se inzeráty nachází.

Nejsou zde popsány všechny HTML značky, které jsou ve funkcích využívány. Všechny HTML značky se nachází ve zdrojovém kódu programu. Na výpisu 6.2 je ukázka funkce *zjistiNazev()*. Z této ukázky je patrné, jak vyhledávání značek funguje.

Ve funkci *informace()* je zjištěn počet inzerátů, makléřů a poboček dané realitní kanceláře. Pokud realitní kancelář neobsahuje ani jeden inzerát, je zpracování přesunuto na další realitní kancelář. Pokud bude mít realitní kancelář alespoň jeden inzerát, program začne získávat a zpracovávat informace o inzerátech realitní

```
#funkce pro zjisteni nazvu realitni kancelare
def zjistiNazev(self,soup):
        #definice proměnné nazev
        nazev = "NA"
        #najdem HTML tag h1 s~třídou page-title
        nadpis = soup.find('h1', attrs={'class' : 'page-title'})
        #pokud tam je pokračujeme, pokud není hodnota je NA
        if nadpis:
                #najdeme textový řetězec, který je uložen ve značce
                     span s~třídou text ng-binding
                nazev = nadpis.find('span', attrs={'class' : 'text
                   ng--binding'}).string
        #uložíme získanou hodnotu do objektu
        self.agencyName = nazev
```

Výpis 6.2: Pomocná funkce pro zjištění názvu realitní kanceláře.

kanceláře. Takové zpracování provádí funkce *zjistiInzerat()*. Této části programu je věnována sekce 6.2.

Nyní je potřeba zjistit zeměpisné souřadnice realitní kanceláře a města, které obsahuje nejvíce inzerátů. Pro zjištění zeměpisných souřadnic sídla realitní kanceláře je vytvořena funkce *getAgencyLongLant()*. Funkce *mostCommon()* zjistí město, které nabízí nejvíce inzerátů. A ve funkci *getNejPocetnejsi()* jsou zjištěny zeměpisné souřadnice a kraj města.

Výsledek se uloží do řetězce s názvem *vypisRk()*. Pokud vše v programu proběhlo bez problému, řetězec se zapíše do souboru *realEstateAgency_time.txt*, kde time je časová značka vytvoření souboru. V tabulce 7.1 je zobrazen příklad výstupního souboru pro sběr realitních kanceláří.

6.1.2 Zpracování informací o realitních kancelářích s parametry

Pokud byl v hlavní funkci programu zadán parametr *--setRealEstateAgencyByName* nebo *--singleRealEstateAgencyByName*, zpracovávání realitních kanceláří funguje na podobném principu jako v předchozí podsekci 6.1.1. Do pole *poleHledanychRk* jsou vložena zadaná jména realitních kanceláří. Po projití všech realitních kanceláří a zjištění jména kanceláře, která se nachází v seznamu *poleHledanychRk*, je vytvořen pro danou kancelář objekt. Pro každý objekt realitní kanceláře je vytvořena složka s jejím ID. Použitím knihoven BeautifulSoup a LXML jsou získány všechny potřebné

údaje pomocí HTML značek na dané stránce. Funkce pro zjištění těchto údajů jsou stejné jako v případě spouštění programu bez parametrů a jsou popsány v 6.1.1. Nakonec je uložena webová stránka v HTML kódu dané realitní kanceláře do souboru a tento soubor se uloží do složky příslušné realitní kanceláře. Pokud vše proběhlo bez potíží, je zapsán řetězec *vypisRk* do souboru *realEstateAgency_time.txt*, kde time je časová značka vytvoření souboru. Nakonec je odstraněn název nalezené realitní kanceláře z pole *poleHledanychRk* a zkontrolována délka pole. Pokud jsou nalezeny všechny zadané realitní kanceláře, délka pole se rovná 0. V tuto dobu je program ukončen, jelikož byly všechny realitní kanceláře úspěšně nalezeny a program by zbytečně procházel další realitní kanceláře. Obsahuje–li realitní kancelář alespoň jeden inzerát, proběhne zpracování inzerátů, které je popsáno v sekci 6.2.

Z polohy nasbíraných inzerátů pro jednotlivé realitní kanceláře je potřeba dále určit město, které má nejvíce inzerátů. Pro toto město je zjištěn i kraj a GPS souřadnice města, dále počet inzerátů v nejčastěji nalezeném městě a počet všech měst, kde jsou inzeráty.

Zjištění města probíhá ve funkci *zjistiInzerat()*. V této funkci je vytvořeno pole s názvem *poleMest*, do kterého budou uložena všechna nalezená města. Funkce *mostCommon()* zjistí nejpočetnější město, dále zjistí i počet inzerátů v nejčastějším městě a počet všech měst, kde jsou inzeráty. Pro zjištění kraje a GPS souřadnic nejčastěji nalezeného města je vytvořena funkce *getNejPocetnejsi()*. Pokud realitní kancelář neobsahuje ani jeden inzerát, je vložena do sloupce pro nejčastější město adresa města, ve kterém má daná realitní kancelář sídlo. Funkce *zjistiWeb()* zjišťuje web realitní kanceláře. Pro položku kontinent bude vždy platit hodnota EU a pro stát vždy hodnota CZ. Nakonec jsou zjištěny zeměpisné souřadnice realitní kanceláře pomocí funkce *getAgencyLongLant()*. Ze zadaného vstupního souboru jsou získány metody.

Pokud byl zadán parametr *--singleRealEstateAgencyByUrl*, musí být současně definována jedna adresa realitní kanceláře z portálu Sreality. Program funguje na podobném principu jako při předchozích parametrech, ale pouze pro jednu webovou stránku inzerátu z portálu Sreality. Program se spustí a pomocí webdriveru Selenium se načte tato jedna webová stránka, z které jsou získány požadované informace.

Pokud bude chtít uživatel zadat parametr *--setRealEstateAgencyByUrl*, spolu s ním se bude muset zadat také parametr *--locationIP*. Pokud parametr *--locationIP* nebude zadán, program se spustí standardním způsobem. Při spuštění programu s parametrem *--setRealEstateAgencyByUrl*, se bude muset definovat i název souboru ve formátu CSV, který obsahuje seznam www serverů realitních kanceláří. V tomto souboru si program vybere pouze ty webové adresy realitních kanceláří, které nejsou hostované a zároveň neobsahují více než 400 inzerátů. Pro tyto realitní kanceláře program provede sběr inzerátů z Sreality bez parametru *--photo*.

Jako první krok při spuštění programu parametrem *--setRealEstateAgencyByUrl*, je otevřen zadaný soubor pro čtení a jsou procházeny všechny www adresy realitních kanceláří. Pokud je některá z nich nehostovaná, je načtena její stránka na portále Sreality. Pomocí knihoven BeautifulSoup a LXML je načtena HTML stránka a začnou se zpracovávat požadované informace. Nejdůležitější informací je počet inzerátů realitní kanceláře. Tato hodnota je zjištěna z funkce *informace()*. Funkce je blíže popsána v podsekci 6.1.1. Pokud nabízí více než 400 inzerátů, je daná realitní kancelář přeskočena. Pokud naopak méně než 400 inzerátu, je vytvořen pro danou realitní kancelář objekt a začnou se zpracovávat požadované informace. Pokud obsahuje realitní kancelář alespoň jeden inzerát, proběhne zpracování inzerátů, které je stejné pro všechny případy zadaných parametrů a je popsáno v sekci 6.2.

Stejně jako u předchozích parametrů, i zde je potřeba zjistit nejpočetnější město, počet inzerátů v tomto městě a počet měst, kde jsou inzeráty. Změna nastane v tom případě, že jsou přidány údaje o ipv4 a ipv6 adresách. Tyto údaje jsou získány ze vstupního souboru.

Z polohy nasbíraných inzerátů pro jednotlivé realitní kanceláře je potřeba určit město, které má nejvíce inzerátů. Pro toto město musí být zjištěn kraj, GPS souřadnice města, počet inzerátů v nejčastěji nalezeném městě a počet všech měst, kde jsou inzeráty. Všechny získané informace jsou vloženy do řetězce *vypisRk*. Pomocí funkce *getAgencyLongLant()* jsou získány zeměpisné souřadnice sídla realitní kanceláře. Nakonec je webová stránka uložena v HTML kódu dané realitní kanceláře do souboru a tento soubor se uloží do složky příslušné realitní kanceláře. Tato složka je vytvořená pro každou realitní kancelář a je pojmenovaná podle unikátního ID realitní kanceláře. Pokud vše proběhlo bez problému, je zapsán řetězec *vypisRk* do souboru *realEstateAgency_time.txt*, kde time je časová značka vytvoření souboru. Formát výsledného souboru se od předešlých zpracování nepatrně liší. Hlavní změnou je to, že byly přidány ipv4a ipv6 adresy nehostovaných realitních kanceláří. Z výsledků programu spuštěného s parametrem *--setRealEstateAgencyByUrl* je tedy možné zjistit přibližnou polohu ipv4 nebo ipv6 adres podle získaného nejpočetnějšího města.

6.2 Získávání a separování dat z webových stránek inzerátů realitních kanceláří

Pokud realitní kancelář obsahuje alespoň jeden inzerát, přesune se do funkce *zjistiInzerat()*. Tato funkce je velice podobná hlavní funkci pro zpracování realitních kanceláří. Procházení seznamu inzerátů a zpracování základních údajů je založeno na podobném principu jako v hlavní funkci. Změna nastává při hledání volitelných

položek inzerátů a vyhledávání adresy a souřadnic inzerátů.

Struktura volitelných položek je zobrazena na obrázku 6.5. Každý inzerát může obsahovat libovolný počet těchto položek. Kdyby se položky zapisovaly postupně tak, jak jsou uloženy na webové stránce, nebyly by seřazeny podle hlavičky souboru a výstupní soubor by potom byl prakticky nečitelný. Proto je vytvořena funkce *zjistiTabulkuInzeratu()*, která zjistí všechny položky na portálu Sreality a uloží je do datového typu slovník. V Pythonu je slovník datový typ, který obsahuje neuspořádané kolekce dvojic klíč-hodnota. Každá položka bude mít svoji hodnotu. Za předpokladu, že hodnotu mít nebude, je uložena hodnota *"NA"*. Poté program projede několik tisíc inzerátů a pokud objeví novou položku, která není ve slovníku, přidá si ji do svého slovníku. Poté je spuštěn program a nalezeny nejčastější položky, které se nacházejí u inzerátů.

Celková cena:	3 990 000 Kč za nemovitost, včetně provize
Poznámka k ceně:	cena již zahrnuje provizi RK, kompletní právní servis s úschovou peněžních prostředků, nezahrnuje daň z nabytí nemovité věci
ID zakázky:	547149
Aktualizace:	Dnes
Stavba:	Cihlová
Stav objektu:	Velmi dobrý
Poloha domu:	Samostatný
Umístění objektu:	Klidná část obce
Typ domu:	Patrový
Podlaží:	3 včetně 1 podzemního
Plocha zastavěná:	107 m²
Užitná plocha:	200 m²
Plocha podlahová:	230 m²

Plocha zahrady:	540 m²
Sklep:	15 m²
Parkování:	3
Rok rekonstrukce:	2012
Voda:	Dálkový vodovod
Topení:	Lokální plynové, Ústřední dálkové
Odpad:	Veřejná kanalizace, Septik
Telekomunikace:	Telefon, Internet
Elektřina:	120V, 230V, 400V
Doprava:	Silnice
Energetická náročnost budovy:	Třída G - Mimořádně nehospodárná č. 148/2007 Sb. podle vyhlášky
Vybavení:	Částečně
Výtah:	✗

Obr. 6.5: Ukázka volitelných položek na v inzerátu realitní kanceláře.

Seznam všech nalezených volitelných položek je zobrazen níže.

- Celková cena
- Hypotéka
- Poznámka k ceně
- ID zakázky
- Aktualizace
- Stavba
- Stav objektu
- Poloha domu
- Umístění objektu

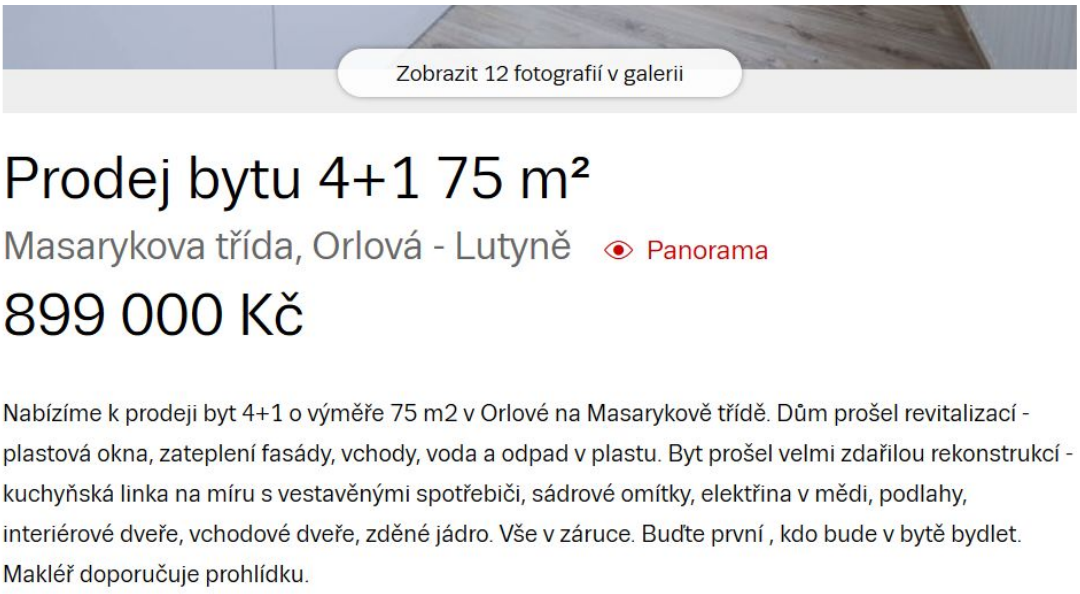
- Typ domu
- Podlaží
- Plocha zastavěná
- Užitná plocha
- Plocha podlahová
- Voda
- Plocha pozemku
- Plocha zahrady
- Sklep
- Parkování
- Topení
- Doprava
- Rok rekonstrukce
- Odpad
- Telekomunikace
- Elektřina
- Energetická náročnost budovy
- Vlastnictví
- Vybavení
- Výtah
- Balkón
- Zlevněno
- Cena za m2
- Původní cena
- Cena
- Garáž
- Plyn
- Minimální kupní cena
- Ukazatel energetické náročnosti budovy
- Komunikace
- Lodžie
- Bezbariérový
- Bazén
- Počet bytů
- Náklady na bydlení
- Převod do OV
- Anuita
- Datum nastěhování
- Rok kolaudace
- Počet kanceláří
- Stav
- ID
- Půdní vestavba
- Výška stropu
- Plocha bazénu
- Vyvolávací cena
- Datum ukončení výstavby
- Provize
- Datum zahájení prodeje
- Znalecký posudek
- Minimální příhoz
- Místo konání dražby
- Aukční jistina
- Druh dražby
- Datum konání dražby
- Počet lůžek
- Termín 1. prohlídky
- Termín 2. prohlídky
- Datum prohlídky
- Terasa

Dále program pracuje s funkcí *zjistiVseInzerat()*. Jejím úkolem je zavolat všechny ostatní pomocné funkce, které jsou potřeba pro získávání informací o inzerátu. Uvnitř této funkce se volají následující funkce:

- *zjistiNazevInzeratu()* – Funkce pro nalezení názvu inzerátu.
- *zjisitAdresuInzeratu()* – Funkce pro nalezení adresy inzerátu.
- *zjisitCenuInzeratu()* – Funkce pro nalezení ceny inzerátu.

- *zjistiMaklereInzeratu()* – Funkce pro nalezení jména a příjmení realitního makléře, příslušícího k danému inzerátu. Další údaje, které jsou zjišťovány o realitním makléři, jsou jeho adresa, mobilní číslo, emailová adresa a webová adresa. Tato funkce je složitější, jelikož jsou HTML značky na portálu Sreality špatně značené a pojmenované. Je nutné, pro nalezení potřebných údajů, použít regulérní výrazy a vyhledávat podle názvu položky.
- *zjisitTextInzeratu()* – Funkce pro nalezení textového popisu příslušící k inzerátu.
- *zjisitTabulkuInzeratu()* – Funkce prohledává seznam položek inzerátů a porovnává ho s položkami, které inzerát skutečně obsahuje. Pokud se v inzerátu daná položka nachází, vloží se do řetězce hodnota této položky, v opačném případě se uloží do řetězce hodnota *"NA"*. Některé položky inzerátu nejsou v textové podobě, ale ve formě ikony. Pokud položka obsahuje ikonu křížku, je ukládána do řetězce hodnota *"NE"* a v případě fajfky hodnota *"ANO"*.

Na obrázku 6.6 je vidět, jak informace o inzerátu, které je potřeba získat, vypadají na serveru Sreality. Na obrázku 6.7 je vidět, jak tato data vypadají v HTML kódu. Pro každý inzerát je vytvořen objekt, a také složka s jejím unikátním identifikátorem. Tato složka bude uložena ve složce realitní kanceláře, ke které inzerát náleží. Všechny zjištěné údaje z předchozích popsaných funkcí jsou ukládány do proměnné *vypisInzeratu*.



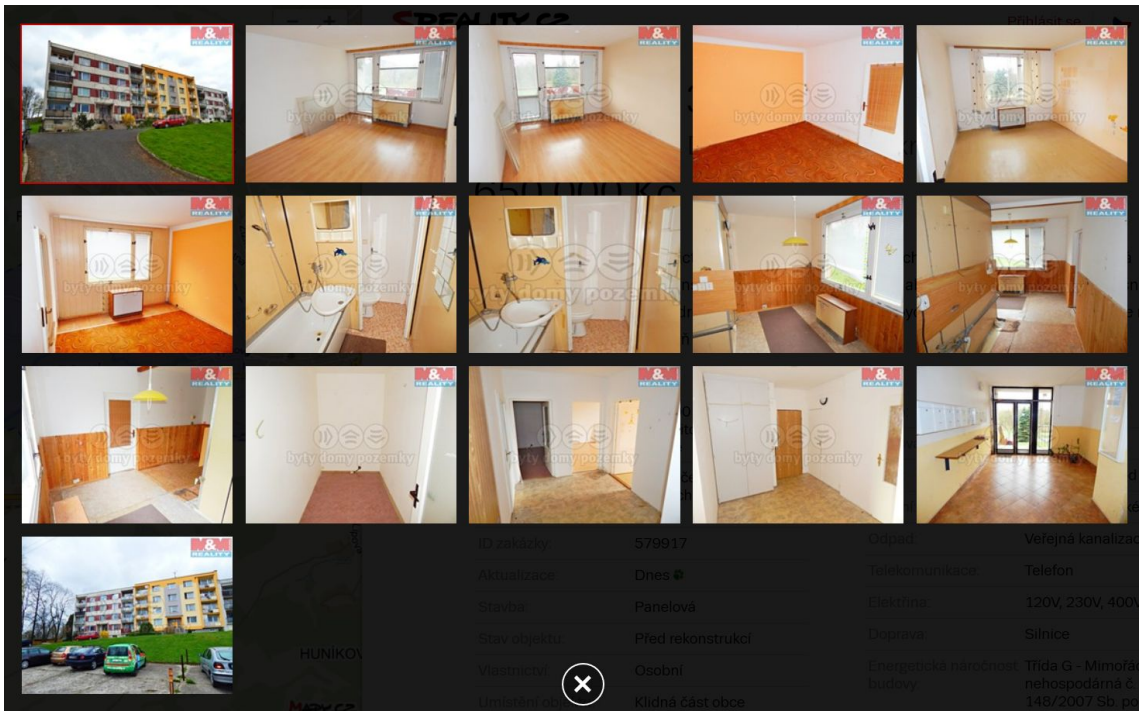
Obr. 6.6: Zobrazení získávaných informací na webu Sreality.

Nyní je potřeba zjistit pro každý inzerát jeho zeměpisnou polohu, ulici, město,

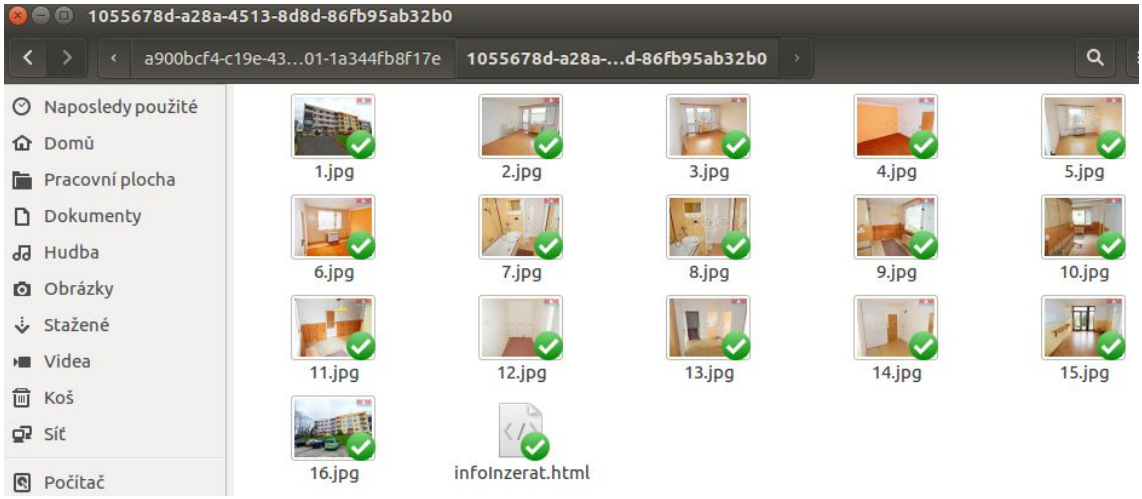
Obr. 6.7: Zobrazení html značek u inzerátu.

okres, kraj a stát. Sekce 6.3 je věnována popisu implementace programu *search.py*, který bude mít za úkol, ze souboru se získanými inzeráty, vyhledávat nemovitosti dle zadané specifikace. K tomuto účelu je vytvořena funkce *getAdress()*, které je v argumentu funkce předávána celá adresa inzerátu. Tato funkce je zařazena do hlavního programu proto, aby program pro vyhledávání nemovitostí *search.py* byl co nejrychlejší. Z této adresy je získána procesem geokódování, za pomocí API od mapy.cz zeměpisná délka, zeměpisná šířka, ulice, město, okres, kraj a stát inzerátu. Tyto údaje jsou uloženy do pole a zapsány na konec příslušného inzerátu. V případě neúspěchu při nalezení údajů ze zadané adresy inzerátu přes API od mapy.cz, je zkoušeno reverzní geokódování, kdy je přes nástroj geopy zjištěna zeměpisná šířka a délka inzerátu a poté nalezeny zbývající údaje. Pokud nejsou nalezena požadovaná data, anebo jsou data ve špatném formátu, vloží se do pole znak *"NA"*. Tato funkce pracuje správně pouze v rámci územně-správních jednotek České republiky. Na konci jsou využity pomocné funkce *uliceParse(), okresParse(), krajParse()*, které odstraní nepotřebné řetězce ze získaných údajů.

V další části funkce je zjišťováno, zda byl zadán parametr *--photo*. Pokud ano, provede se stáhnutí všech fotek, které daný inzerát obsahuje. Na obrázku 6.8 je vidět, jak se fotky zobrazují na serveru Sreality. Do složky jsou ukládány pouze náhledy fotek, zejména kvůli velikosti celých fotek. Stahováním náhledu fotek je dotazováno pouze na jednu webovou stránku, ale kdyby se program pokoušel získat fotky v jejich normální velikosti, musel by využít o mnoho více dotazů na webové stránky, což by zpomalovalo běh programu. Všechny fotky daného inzerátu se poté uloží do složky s ID příslušného inzerátu. Jak vypadají tyto fotky ve složce, je možné vidět na obrázku 6.9. Do této složky se vloží i stránka inzerátu ve formátu HTML.



Obr. 6.8: Zobrazení detailu fotek na portálu Sreality.



Obr. 6.9: Zobrazení fotek v složce.

Nakonec je potřeba zjistit údaje jako nejpočetnější město, počet inzerátu v tomto městě, počet měst, kde jsou inzeráty a zeměpisné souřadnice nejpočetnějšího města. Ve funkci *zjistiInzerat()* se vytvoří pole s názvem *poleMest*, do kterého se budou ukládat všechna nalezená města. Funkce *mostCommon()* zjistí nejpočetnější město. V další části funkce se zjistí počet inzerátů v nejčastějším městě a počet všech měst,

kde jsou inzeráty. Pro zjištění kraje a GPS souřadnic nejčastěji nalezeného města je vytvořena funkce *getNejPocetnejsi()*. Pokud realitní kancelář neobsahuje ani jeden inzerát, je vložena do sloupce pro nejčastější město, adresa města, ve kterém má daná realitní kancelář sídlo. Pro položku kontinent bude vždy platit hodnota EU a pro stát vždy hodnota CZ. Zbytek potřebných údajů jako jsou metody, jsou získány ze zadaného vstupního souboru. V tabulce 7.5 je zobrazena ukázka výstupního souboru pro získávání údajů o inzerátech.

Nakonec jsou všechny informace o inzerátu uloženy do textové proměnné *vypisVsechInzeratu*. Pokud program proběhne v pořádku, zapíší se získané informace do souboru *inzeraty_time.txt*, kde time je časová značka vytvoření souboru.

6.3 Program pro vyhledávání nemovitostí podle zadaných parametrů

V této části se provede popis implementace programu *search.py*, který bude mít za úkol najít vhodné nemovitosti podle zadaných kritérií. Program je možné spouštět s následujícími parametry:

- *-h* nebo *--help* – Tento parametr vypíše nápovědu k programu. Může být zadán pouze jednou a nesmí se kombinovat s žádným jiným parametrem nebo skončí s chybou.
- *--input* – Tento parametr specifikuje název složky, ve které bude probíhat vyhledávání souborů s uloženými inzeráty.
- *--output* – Tento parametr definuje název výstupního souboru, do kterého se budou ukládat výsledky. Pokud tento parametr není zadán, výsledek se vypíše na standardní výstup.
- *--fromDate* – Tento parametr určuje časový úsek, od kterého má program začít vyhledávat nemovitosti. Tento parametr se zadává ve formátu ROK–MĚSÍC (např. 2017-03).
- *--toDate* – Tento parametr určuje časový úsek, do kterého má program vyhledávat nemovitosti. Tento parametr se zadává ve formátu ROK–MĚSÍC (např. 2017-03).
- *--street* – Tento parametr specifikuje hledanou ulici, na které se nachází nemovitost.
- *--city* – Tento parametr specifikuje hledané město nebo vesnici, ve kterém bude program hledat nemovitosti.
- *--district* – Tento parametr specifikuje hledaný okres nemovitosti.
- *--region* – Tento parametr specifikuje hledaný kraj nemovitosti.
- *--country* – Tento parametr specifikuje hledaný stát nemovitosti.

- *--type* – Tento parametr specifikuje hledaný typ nemovitosti. Inzerát může obsahovat typy jako prodej, pronájem nebo dražba.
- *--minPrice* – Tento parametr specifikuje minimální cenu nemovitosti.
- *--maxPrice* – Tento parametr specifikuje maximální cenu nemovitosti.
- *--distance* – Tento parametr specifikuje vzdálenost nemovitostí od zadaného města nebo vesnice. Jednotka délky, se kterou parametr pracuje, je v kilometrech. Tento parametr musí být zadán zároveň s parametrem *--city*. Pro upřesnění vzdálenosti je možno využít k parametru *--city* i parametr *--street*.
- *--keyword* – Tento parametr specifikuje, které položky nemovitostí chce uživatel hledat (příklad: *--keyword="Sklep"*, vyhledá všechny inzeráty, které vlastní sklep). Program dokáže vyhledávat položky nemovitostí i reverzně. Pro reverzní vyhledávání je potřeba zadat před hledané slovo znaménko "–"(příklad: *--keyword="–Parkování"*, vyhledá všechny inzeráty, které nemají parkování). Seznam položek, které lze v tomto parametru využít, je popsaný výše.

Všechny parametry, kromě parametru *--keyword*, musí být zadané pouze jednou.

Nyní je popsáno, na jakém principu program pracuje. V hlavní funkci programu je otevřena zadaná složka a porovnávány názvy jejich podsložek, zda se rovnají časové době, kterou zadal uživatel v parametrech příkazového řádku. Pokud žádný časový údaj uživatel nezadal, prohledávají se všechny podsložky. Pokud se časový údaj podsložky shoduje s časovým rozmezím zadaného uživatelem, program může začít s danou podsložkou pracovat. Když se program dostane do zadané podsložky, otevře si textový soubor, který ve svém názvu obsahuje řetězec *"inzeraty"*. Poté program zavolá funkci *filtrujSoubor()*, jehož argumenty jsou prohledávaný soubor s inzeráty, parametry příkazové řádku zadané uživatelem a proměnná, která určuje, kam zapsat výsledky. Funkce *filtrujSoubor()* nejdříve zjistí, zda byla zadaná klíčová slova. Pokud ano, je přečten první řádek souboru, na kterém se nachází uložená hlavička souboru. Dále je zjištěno, zda dané klíčové slovo existuje. Pokud se zadané klíčové slovo nachází v hlavičce souboru, je potřeba zjistit, zda se jedná o nereverzní či reverzní klíčové slovo. Reverzní klíčové slovo má před hledaným klíčovým slovem znak *"–"* a vyhledá ty inzeráty, které obsahují v hledaném klíčovém slově řetězec *"NA"* nebo *"NE"*. Příklad reverzního vyhledávání: *--keyword="-Parkování"*. Pokud se slovo v hlavičce shoduje se slovem klíčovým, je vytvořen objekt *klicoveSlovo*, do kterého je vloženo nalezené klíčové slovo, index sloupce, ve kterém je klíčové slovo uloženo a příznak, zda se jedná o nereverzní či reverzní klíčové slovo. Nakonec je objekt vložen do pole s názvem *polePozicKeywords*.

Další možností zpracování vyhledávání inzerátu je podle zadané vzdálenosti. Pokud chce uživatel použít parametr pro určení vzdálenosti, musí být současně zadán i parametr pro město, od jehož vzdálenosti jsou inzeráty vyhledávány. Pokud by uživatel chtěl, může upřesnit lokalitu i zadáním parametru ulice. Ve funkci *getAdress()*

s argumenty ulice a město, je zjištěna zeměpisná šířka a délka uživatelem zadané lokality. K získání souřadnic lokality je využíváno API pro reverzní geokódování od mapy.cz. Pokud tato API nic nenajde, je využita knihovna geopy pro získání souřadnic lokality. Po získání zeměpisných souřadnic hledané lokality je znovu využita knihovna geopy. Tato knihovna dokáže spočítat zeměpisnou vzdálenost mezi dvěma body. Využívá algoritmy jako Vincentyho formule nebo Great Circle. Pro získání této vzdálenosti je dostupná funkce *geopy.distance.distance()*. Poté je porovnáváno, jestli získaná vzdálenost souhlasí s požadavkem uživatele. Pokud ne, tak není nemovitost zahrnuta do dalších filtrů. Ve výpisu 6.3 je ukázka získání zeměpisné vzdálenosti mezi dvěma body.

```
#bod p1 = souřadnice získané hledaním lokality podle zadání už
    ivatele
p1 = geopy.Point(float(adresaPole[1]), float(adresaPole[0]))
#bod p2 = souřadnice našich inzerátů
p2 = geopy.Point(float(sloupce[89]), float(sloupce[88]))
#výsledke v~kilometrech
result = geopy.distance.distance(p1,p2).km

        #porovnávání zda výsledek splňuje podmínku
        if(float(result) <= float(vzdalenost)):
                #nastavení příznaku
                veVzdalenosti = True
```

Výpis 6.3: Využití funkce *geopy.distance.distance* pro získání zeměpisné vzdálenosti mezi dvěma body.

Aplikace vyhledávacího filtru porovnává parametry zadané od uživatele s daty v prohledávaném textovém souboru. Výhodou je, že vstupní soubor je uložený ve sloupcích, tudíž jsou porovnávány zadané hodnoty od uživatele, s hodnotami uloženými ve sloupcích prohledávaného souboru. Pokud inzerát vyhovuje specifikaci uživatele, je filtrován podle zbývajících parametrů jako cena a klíčová slova. Některé inzeráty ceny neobsahují, místo toho sdělují, že informaci o ceně dostaneme u dané realitní kanceláře. Tyto inzeráty jsou do výsledného souboru také zahrnuty. Pokud se najde nemovitost, která splní všechny podmínky, je zapsána do souboru nebo na standardní výstup.

7 DEMONSTRACE ZÍSKANÝCH DAT A JEJICH APLIKACE

V této kapitole jsou demonstrovány výsledky aplikací. Je představena výsledná struktura výstupních souborů a příklady spuštění programu s různými parametry a jejich výsledky. Příklady spuštění aplikací v této kapitole:

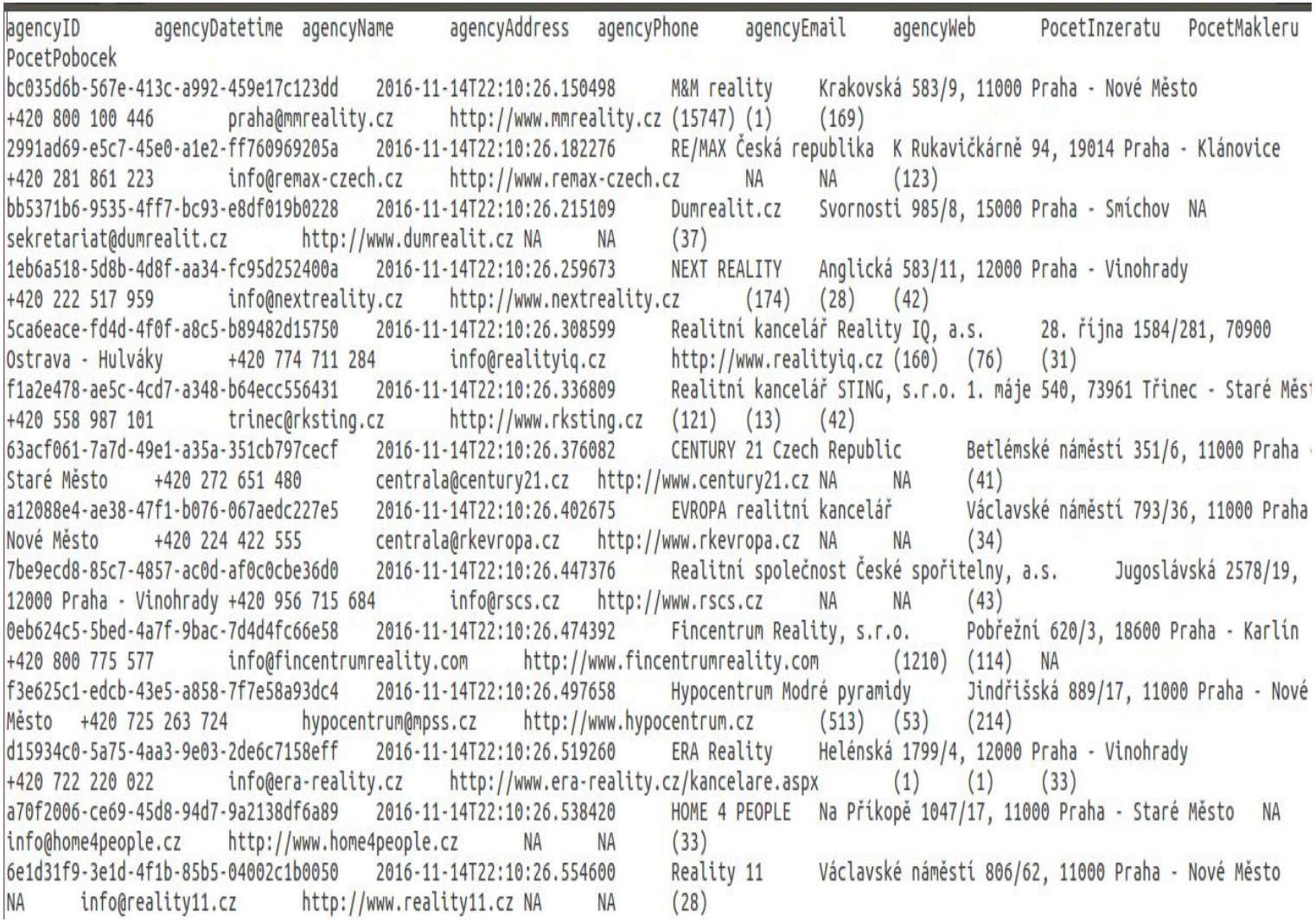
- Příklad na získání údajů všech realitních kanceláří.
- Příklad na získání inzerátu největší realitní kanceláře podle její webové adresy (M&M reality).
- Příklad na získání inzerátu jedné realitní kanceláře, podle jejího jména (REMAX High Way Kolín).
- Příklady pro nalezení nemovitostí se zvolenými parametry.

Aplikace je naprogramovaná v jazyce Python verze 3.5.2 a pracuje v 64 bitovém operačním systému Ubuntu 16.04 LTS. Aplikace byla otestována na webovém prohlížeči Mozilla Firefox verze 45.0. Selenium webdriver pracuje na verzi 2.53.6. Verze knihovny BeautifulSoup je 4.4.0. Operační paměť počítače, na kterém aplikace běží, je 3.7 GiB s procesorem Intel Core i3 CPU M 330 2.13 GHz x 4. Rychlost připojení byla 50 Mb/s, sítě VUT.

Výstupem aplikace pro zpracování veřejně dostupných dat z internetu je složka, která se nazývá *output_time*, kde time je časová značka začátku programu. V této složce se nachází dva textové soubory.

První textový soubor je pojmenovaný *realEstateAgency_time.txt*, kde time je časová značka vytvoření souboru. V tomto souboru jsou uložena získaná a separovaná data o realitních kancelářích, uložených na portálu Sreality. Program byl několikrát testován a zpracoval průměrně 2600 realitních kanceláří. Průměrně proto, že malé realitní kanceláře se na portálu rychle objevují, ale i rychle mizí. Jedná se hlavně o realitní kanceláře s jedním nebo dvěma inzeráty. Velké realitní kanceláře se příliš často nemění. Velikost tohoto souboru je průměrně 650 Kb. Sběr informací o realitních kancelářích bez inzerátů a fotek trvá přibližně 4 a půl hodiny. Tyto údaje jsou shrnuty v tabulce 7.6. Textový výstup souboru pro realitní kanceláře je možné vidět na obrázku 7.1. Díky tomu, že je v návěští každá položka oddělená tabulátorem, je textový výstup seřazen do sloupců, například pomocí programu LibreOffice. Na obrázku 7.2 je zobrazena ukázka výstupního souboru, seřazeného do sloupců v programu LibreOffice. V tabulce 7.1 je zobrazena ukázka výstupního souboru pro získávání údajů o realitních kancelářích.

Prvním příkladem je spuštění aplikace pro nalezení informací o všech realitních kancelářích. Program byl spuštěn následujícím příkazem 7.1. Shrnutí použitých parametrů příkazu se nachází v tabulce 7.2. Na obrázku 7.3 jsou na mapě vyznačeny



```
agencyID	agencyDatetime	agencyName	agencyAddress	agencyPhone	agencyEmail	agencyWeb	PocetInzeratu	PocetMakleru
PocetPobocek
bc035d6b-567e-413c-a992-459e17c123dd	2016-11-14T22:10:26.150498	M&M reality	Krakovská 583/9, 11000 Praha - Nové Město
+420 800 100 446	praha@mmreality.cz	http://www.mmreality.cz (15747) (1)	(169)
2991ad69-e5c7-45e0-a1e2-ff760969205a	2016-11-14T22:10:26.182276	RE/MAX Česká republika	K Rukavičkárně 94, 19014 Praha - Klánovice
+420 281 861 223	info@remax-czech.cz	http://www.remax-czech.cz	NA	NA	(123)
bb5371b6-9535-4ff7-bc93-e8df019b0228	2016-11-14T22:10:26.215109	Dumrealit.cz	Svornosti 985/8, 15000 Praha - Smíchov	NA
sekretariat@dumrealit.cz	http://www.dumrealit.cz NA	NA	(37)
1eb6a518-5d8b-4d8f-aa34-fc95d252400a	2016-11-14T22:10:26.259673	NEXT REALITY	Anglická 583/11, 12000 Praha - Vinohrady
+420 222 517 959	info@nextreality.cz	http://www.nextreality.cz	(174)	(28)	(42)
5ca6eace-fd4d-4f0f-a8c5-b89482d15750	2016-11-14T22:10:26.308599	Realitní kancelář Reality IQ, a.s.	28. října 1584/281, 70900
Ostrava - Hulváky	+420 774 711 284	info@realityiq.cz	http://www.realityiq.cz (160)	(76)	(31)
f1a2e478-ae5c-4cd7-a348-b64ecc556431	2016-11-14T22:10:26.336809	Realitní kancelář STING, s.r.o. 1. máje 540, 73961 Třinec - Staré Měs
+420 558 987 101	trinec@rksting.cz	http://www.rksting.cz	(121)	(13)	(42)
63acf061-7a7d-49e1-a35a-351cb797cecf	2016-11-14T22:10:26.376082	CENTURY 21 Czech Republic	Betlémské náměstí 351/6, 11000 Praha
Staré Město	+420 272 651 480	centrala@century21.cz	http://www.century21.cz NA	NA	(41)
a12088e4-ae38-47f1-b076-067aedc227e5	2016-11-14T22:10:26.402675	EVROPA realitní kancelář	Václavské náměstí 793/36, 11000 Praha
Nové Město	+420 224 422 555	centrala@rkevropa.cz	http://www.rkevropa.cz	NA	NA	(34)
7be9ecd8-85c7-4857-ac0d-af0c0cbe36d0	2016-11-14T22:10:26.447376	Realitní společnost České spořitelny, a.s.	Jugoslávská 2578/19,
12000 Praha - Vinohrady +420 956 715 684	info@rscs.cz	http://www.rscs.cz	NA	NA	(43)
0eb624c5-5bed-4a7f-9bac-7d4d4fc66e58	2016-11-14T22:10:26.474392	Fincentrum Reality, s.r.o.	Pobřežní 620/3, 18600 Praha - Karlín
+420 800 775 577	info@fincentrumreality.com	http://www.fincentrumreality.com	(1210)	(114)	NA
f3e625c1-edcb-43e5-a858-7f7e58a93dc4	2016-11-14T22:10:26.497658	Hypocentrum Modré pyramidy	Jindřišská 889/17, 11000 Praha - Nové
Město	+420 725 263 724	hypocentrum@mpss.cz	http://www.hypocentrum.cz	(513)	(53)	(214)
d15934c0-5a75-4aa3-9e03-2de6c7158eff	2016-11-14T22:10:26.519260	ERA Reality	Helénská 1799/4, 12000 Praha - Vinohrady
+420 722 220 022	info@era-reality.cz	http://www.era-reality.cz/kancelare.aspx	(1)	(1)	(33)
a70f2006-ce69-45d8-94d7-9a2138df6a89	2016-11-14T22:10:26.538420	HOME 4 PEOPLE	Na Příkopě 1047/17, 11000 Praha - Staré Město	NA
info@home4people.cz	http://www.home4people.cz	NA	NA	(33)
6e1d31f9-3e1d-4f1b-85b5-04002c1b0050	2016-11-14T22:10:26.554600	Reality 11	Václavské náměstí 806/62, 11000 Praha - Nové Město
NA	info@reality11.cz	http://www.reality11.cz NA	NA	(28)
```

Obr. 7.1: Ukázka výstupního textového souboru pro realitní kanceláře.

agencyName	agencyAddress	agencyPhone	agencyEmail	agencyWeb	PocetInzerat	PocetMakler	PocetPobocek
M&M reality	Krakovská 583/9, 11000 Praha - Nové Město	420800100446	praha@mmreality.cz	http://www.mmreality.cz	(15747)	(1)	(169)
RE/MAX Česká republika	K Rukavičkárně 94, 19014 Praha - Klánovice	420281861223	info@remax-czech.cz	http://www.remax-czech.cz	NA	NA	(123)
Dumrealit.cz	Svornosti 985/8, 15000 Praha - Smíchov	NA	sekretariat@dumrealit.cz	http://www.dumrealit.cz	NA	NA	(37)
NEXT REALITY	Anglická 583/11, 12000 Praha - Vinohrady	420222517959	info@nextreality.cz	http://www.nextreality.cz	(174)	(28)	(42)
Realitní kancelář Reality IQ, a.s.	28. října 1584/281, 70900 Ostrava - Hulváky	420774711284	info@realityiq.cz	http://www.realityiq.cz	(160)	(76)	(31)
Realitní kancelář STING, s.r.o.	1. máje 540, 73961 Třinec - Staré Město	420558987101	trinec@rksting.cz	http://www.rksting.cz	(121)	(13)	(42)
CENTURY 21 Czech Republic	Betlémské náměstí 351/6, 11000 Praha - Staré Mě	420272651480	centrala@century21.cz	http://www.century21.cz	NA	NA	(41)
EVROPA realitní kancelář	Václavské náměstí 793/36, 11000 Praha - Nové M	420224422555	centrala@rkevropa.cz	http://www.rkevropa.cz	NA	NA	(34)
Realitní společnost České spořitel	Jugoslávská 2578/19, 12000 Praha - Vinohrady	420956715684	info@rscs.cz	http://www.rscs.cz	NA	NA	(43)
Fincentrum Reality, s.r.o.	Pobřežní 620/3, 18600 Praha - Karlín	420800775577	info@fincentrumreality.com	http://www.fincentrumreality.c	(1210)	(114)	NA
Hypocentrum Modré pyramidy	Jindřišská 889/17, 11000 Praha - Nové Město	420725263724	hypocentrum@mpss.cz	http://www.hypocentrum.cz	(513)	(53)	(214)
ERA Reality	Helénská 1799/4, 12000 Praha - Vinohrady	420722220022	info@era-reality.cz	http://www.era-reality.cz/kanc	(1)	(1)	(33)
HOME 4 PEOPLE	Na Příkopě 1047/17, 11000 Praha - Staré Město	NA	info@home4people.cz	http://www.home4people.cz	NA	NA	(33)
Reality 11	Václavské náměstí 806/62, 11000 Praha - Nové M	NA	info@reality11.cz	http://www.reality11.cz	NA	NA	(28)
RPG Byty, s.r.o.	Gregorova 2582/3, 70200 Ostrava - Moravská Ostr	420840114115	pronajmybytu@rpgbyty.cz	http://www.rpgbyty.cz	(815)	(26)	(2)
108 AGENCY, s.r.o.	Přibenická 939/20, 13000 Praha - Žižkov	420222211464	info@108agency.cz	http://www.108agency.cz	(659)	(12)	NA
COLOSEUM NEMOVITOSTI s.r.o.	Květná 167/3, 60300 Brno - Pisárky	420800536536	info@coloseumreality.cz	http://www.coloseumreality.cz	(70)	(11)	(7)
SVOBODA & WILLIAMS, s.r.o.	Na Perštýně 362/2, 11000 Praha - Staré Město	420257328281	info@svoboda-williams.com	http://www.svoboda-williams.c	(577)	(1)	(2)

Obr. 7.2: Ukázka textového souboru pro realitní kanceláře se seřazenými položkami v aplikaci LibreOffice.

agencyID	fed9e356-8d65-406a-bf01-b2ff7d300f20
agencyDatetime	2017-04-17T16:55:48
agencyName	M&M reality
agencyAddress	Krakovská 583/9, 11000 Praha - Nové Město
agencyPhone	+420 800 100 446
agencyEmail	praha@mmreality.cz
agencyWeb	http://www.mmreality.cz
PocetMakleru	60
PocetPobocek	166
SrealityWeb	https://www.sreality.cz/adresar/mm-reality-praha-nove-mesto/1019
Kontinent	EU
Stát	CZ
Kraj	Moravskoslezský
Město	Ostrava
Latitude Real Estate	49.8349139
Longitude Real Estate	18.2820084
Latitude Real Estate Agency	50.0780638673
Longitude Real Estate Agency	14.4277479692
Počet inzerátů	15003
Počet inzerátů ve městě	842
Počet měst kde jsou inzeráty	2767
Metoda 1	NA
Metoda 2	NA
Metoda 3	NA
Metoda 4	NA
Hostováno	NA

Tab. 7.1: Ukázka výstupního souboru pro sběr informací o realitních kanceláří.

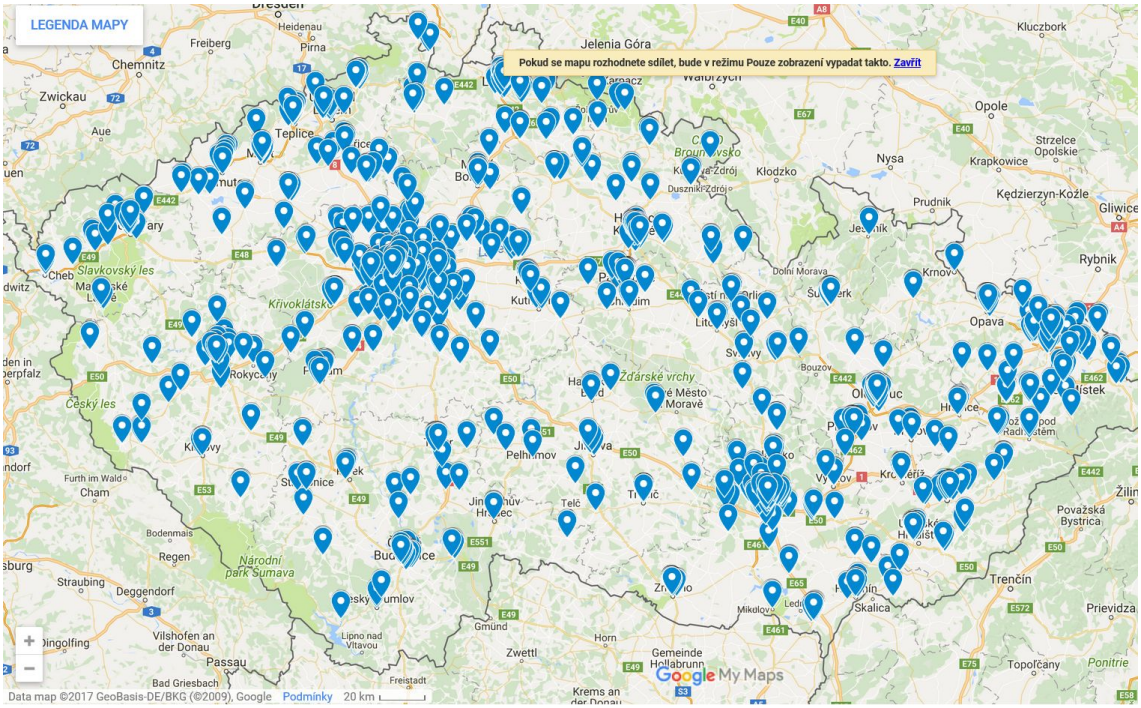
získané polohy realitních kanceláří.

```
$ python3 diplomka.py --output . --realEstate 0
```

Výpis 7.1: Spuštění programu pro získání dat o realitních kancelářích.

--output	.
--realEstate	0

Tab. 7.2: Parametry pro sbíraní informací o realitních kancelářích, bez inzerátů.



Obr. 7.3: Zobrazení poloh realitních kanceláří.

Pokud byl zadán parametr *--setRealEstateAgencyByUrl*, načte se CSV soubor, ve kterém je seznam realitních kanceláří a jejich IP adres. Program zjistí polohy inzerátů a nejčastější město pro inzeráty. Z těchto údajů se zjistí poloha IP adres. Velikost výsledného souboru a čas běhu programu záleží na počtu nehostovaných realitních kanceláří. Tento soubor má velmi podobnou strukturu jako předchozí soubor o realitních kancelářích, ukázaný v tabulce 7.1.

Druhý textový soubor má název *inzeraty_time.txt*, kde time je časová značka vytvoření souboru. V tomto souboru se nachází sesbíraná data všech inzerátů. Průměrný počet inzerátů je přibližně okolo 95 tisíc. Tento soubor má velikost přibližně

140 MB. Pro každý inzerát se program pokoušel najít i adresy pro pozdější vyhledávání inzerátů podle lokality. K získávání poloh inzerátu byla využita API od mapy.cz. Pokud API adresu nenašlo, proběhlo hledání polohy reverzně pomocí knihovny geopy. Nalezení polohy inzerátů proběhlo s úspěšností přibližně 99.5%. Textový výstup souboru pro inzeráty realitní kanceláře je možné vidět na obrázku 7.4. V tabulce 7.5 je zobrazena ukázka výstupního souboru pro získávání údajů o inzerátech.

Na výpisu 7.2 je ukázáno spuštění programu pro největší realitní kancelář na serveru Sreality. Shrnutí použitých parametrů příkazu se nachází v tabulce 7.3. Tato realitní kancelář se nazývá M&M reality. Aplikace sesbírá data o inzerátech, které tato realitní kancelář nabízí.

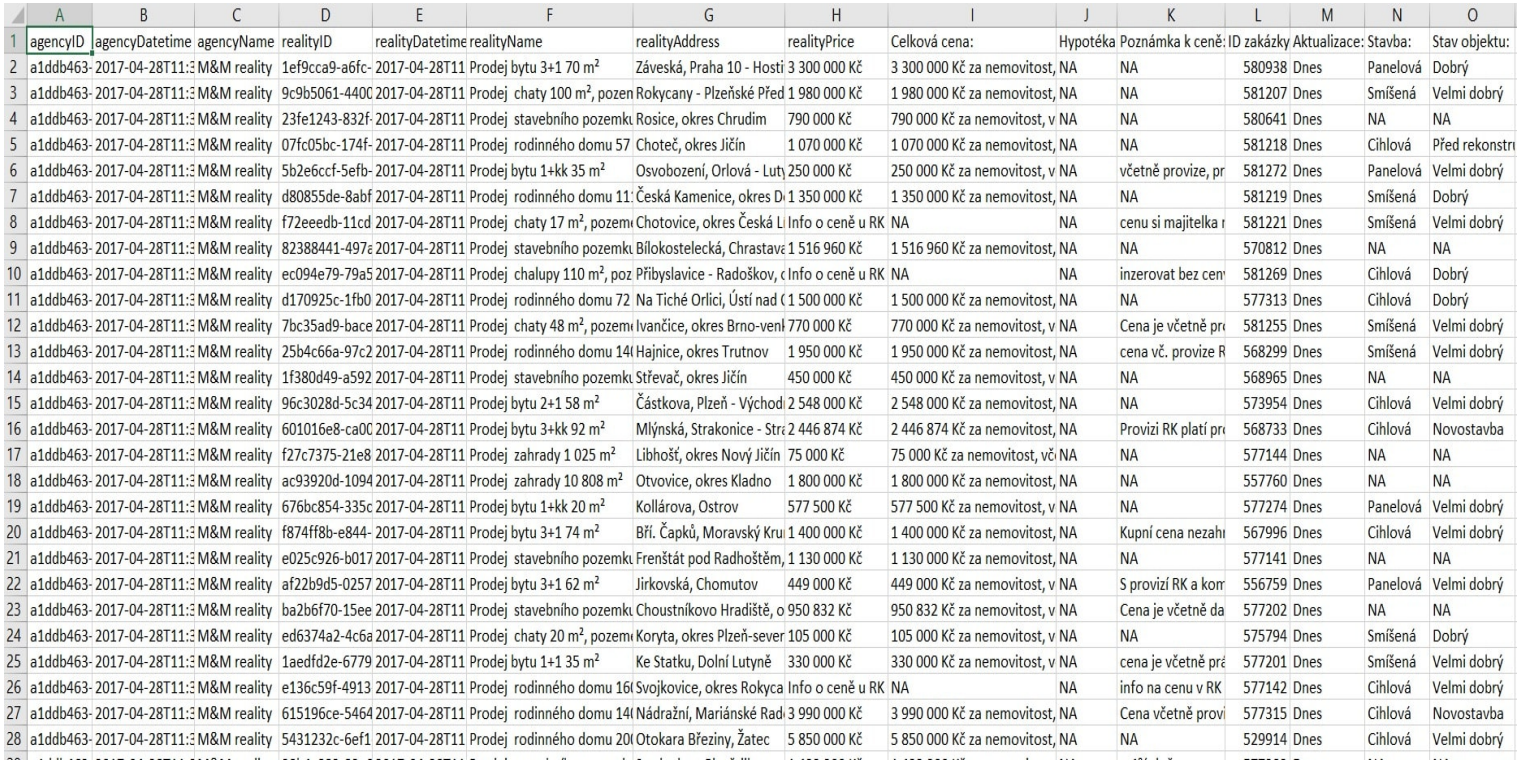
```
$ python3 diplomka.py --output . --singleRealEstateAgencyByUrl="https://
    www.sreality.cz/adresar/mm-reality-praha-nove-mesto/1019"
```

Výpis 7.2: Spuštění programu pro sbírání informací o M&M reality.

--output	.
--singleRealEstateAgencyByUrl	https://www.sreality.cz/adresar/mm-reality-praha-nove-mesto/1019

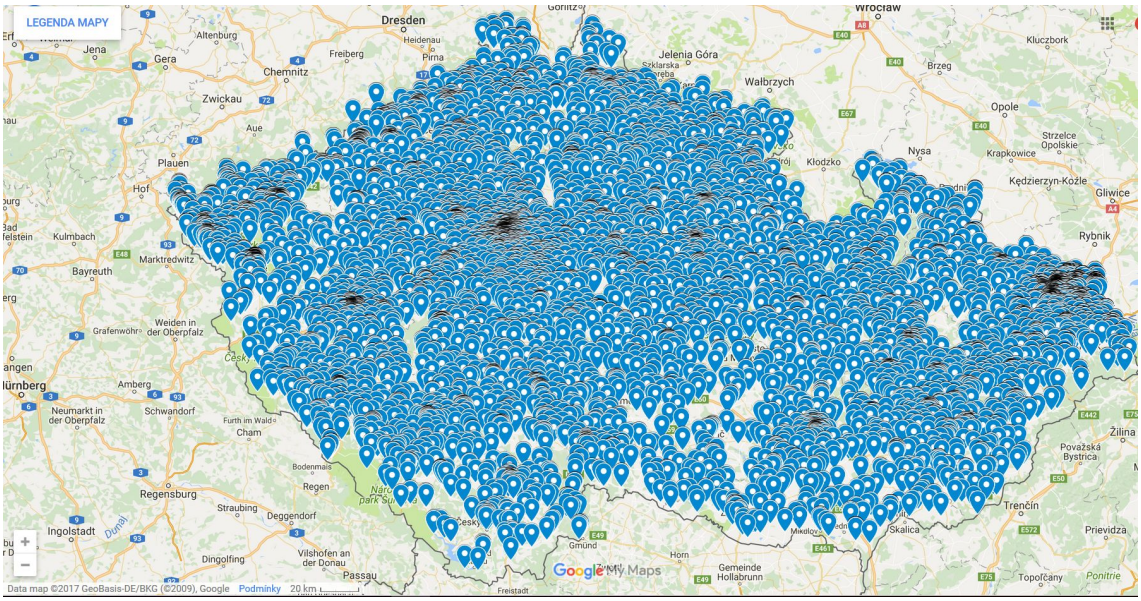
Tab. 7.3: Parametry pro sbírání informací o realitní kanceláři M&M reality a inzerátech, které nabízí.

Na obrázku jsou zobrazeny zjištěné polohy všech inzerátů realitní kanceláře M&M reality. Počet těchto inzerátů se pohyboval okolo 15 000. Z obrázku 7.5 je vidět, že tato realitní kancelář nabízí inzeráty po celé České republice.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	agencyID	agencyDatetime	agencyName	realityID	realityDatetime	realityName	realityAddress	realityPrice	Celková cena:	Hypotéka	Poznámka k ceně:	ID zakázky	Aktualizace:	Stavba:	Stav objektu:
2	a1ddb463-	2017-04-28T11:3	M&M reality	1ef9cca9-a6fc-	2017-04-28T11	Prodej bytu 3+1 70 m²	Záveská, Praha 10 - Hosti	3 300 000 Kč	3 300 000 Kč za nemovitost,	NA	NA	580938	Dnes	Panelová	Dobrý
3	a1ddb463-	2017-04-28T11:3	M&M reality	9c9b5061-4400	2017-04-28T11	Prodej chaty 100 m², pozen	Rokycany - Plzeňské Před	1 980 000 Kč	1 980 000 Kč za nemovitost,	NA	NA	581207	Dnes	Smíšená	Velmi dobrý
4	a1ddb463-	2017-04-28T11:3	M&M reality	23fe1243-832f-	2017-04-28T11	Prodej stavebního pozemku	Rosice, okres Chrudim	790 000 Kč	790 000 Kč za nemovitost, v	NA	NA	580641	Dnes	NA	NA
5	a1ddb463-	2017-04-28T11:3	M&M reality	07fc05bc-174f-	2017-04-28T11	Prodej rodinného domu 57	Choteč, okres Jičín	1 070 000 Kč	1 070 000 Kč za nemovitost,	NA	NA	581218	Dnes	Cihlová	Před rekonstr
6	a1ddb463-	2017-04-28T11:3	M&M reality	5b2e6ccf-5efb-	2017-04-28T11	Prodej bytu 1+kk 35 m²	Osvobození, Orlová - Lut	250 000 Kč	250 000 Kč za nemovitost, v	NA	včetně provize, pr	581272	Dnes	Panelová	Velmi dobrý
7	a1ddb463-	2017-04-28T11:3	M&M reality	d80855de-8abf	2017-04-28T11	Prodej rodinného domu 11	Česká Kamenice, okres D	1 350 000 Kč	1 350 000 Kč za nemovitost,	NA	NA	581219	Dnes	Smíšená	Dobrý
8	a1ddb463-	2017-04-28T11:3	M&M reality	f72eeedb-11cd	2017-04-28T11	Prodej chaty 17 m², pozem	Chotovice, okres Česká L	Info o ceně u RK	NA	NA	cenu si majitelka r	581221	Dnes	Smíšená	Velmi dobrý
9	a1ddb463-	2017-04-28T11:3	M&M reality	82388441-497a	2017-04-28T11	Prodej stavebního pozemku	Bílokostelecká, Chrastava	1 516 960 Kč	1 516 960 Kč za nemovitost,	NA	NA	570812	Dnes	NA	NA
10	a1ddb463-	2017-04-28T11:3	M&M reality	ec094e79-79a5	2017-04-28T11	Prodej chalupy 110 m², poz	Přibyslavice - Radoškov, (	Info o ceně u RK	NA	NA	inzerovat bez cen	581269	Dnes	Cihlová	Dobrý
11	a1ddb463-	2017-04-28T11:3	M&M reality	d170925c-1fb0	2017-04-28T11	Prodej rodinného domu 72	Na Tiché Orlici, Ústí nad (	1 500 000 Kč	1 500 000 Kč za nemovitost,	NA	NA	577313	Dnes	Cihlová	Dobrý
12	a1ddb463-	2017-04-28T11:3	M&M reality	7bc35ad9-bace	2017-04-28T11	Prodej chaty 48 m², pozem	Ivančice, okres Brno-ven	770 000 Kč	770 000 Kč za nemovitost, v	NA	Cena je včetně pr	581255	Dnes	Smíšená	Velmi dobrý
13	a1ddb463-	2017-04-28T11:3	M&M reality	25b4c66a-97c2	2017-04-28T11	Prodej rodinného domu 14	Hajnice, okres Trutnov	1 950 000 Kč	1 950 000 Kč za nemovitost,	NA	cena vč. provize R	568299	Dnes	Smíšená	Velmi dobrý
14	a1ddb463-	2017-04-28T11:3	M&M reality	1f380d49-a592	2017-04-28T11	Prodej stavebního pozemku	Střevač, okres Jičín	450 000 Kč	450 000 Kč za nemovitost, v	NA	NA	568965	Dnes	NA	NA
15	a1ddb463-	2017-04-28T11:3	M&M reality	96c3028d-5c34	2017-04-28T11	Prodej bytu 2+1 58 m²	Částkova, Plzeň - Východ	2 548 000 Kč	2 548 000 Kč za nemovitost,	NA	NA	573954	Dnes	Cihlová	Velmi dobrý
16	a1ddb463-	2017-04-28T11:3	M&M reality	601016e8-ca0(	2017-04-28T11	Prodej bytu 3+kk 92 m²	Mlýnská, Strakonice - Stra	2 446 874 Kč	2 446 874 Kč za nemovitost,	NA	Provizi RK platí pr	568733	Dnes	Cihlová	Novostavba
17	a1ddb463-	2017-04-28T11:3	M&M reality	f27c7375-21e8	2017-04-28T11	Prodej zahrady 1 025 m²	Libhošť, okres Nový Jičín	75 000 Kč	75 000 Kč za nemovitost, vč	NA	NA	577144	Dnes	NA	NA
18	a1ddb463-	2017-04-28T11:3	M&M reality	ac93920d-1094	2017-04-28T11	Prodej zahrady 10 808 m²	Otvovice, okres Kladno	1 800 000 Kč	1 800 000 Kč za nemovitost,	NA	NA	557760	Dnes	NA	NA
19	a1ddb463-	2017-04-28T11:3	M&M reality	676bc854-335c	2017-04-28T11	Prodej bytu 1+kk 20 m²	Kollárova, Ostrov	577 500 Kč	577 500 Kč za nemovitost, v	NA	NA	577274	Dnes	Panelová	Velmi dobrý
20	a1ddb463-	2017-04-28T11:3	M&M reality	f874ff8b-e844-	2017-04-28T11	Prodej bytu 3+1 74 m²	Bří. Čapků, Moravský Kru	1 400 000 Kč	1 400 000 Kč za nemovitost,	NA	Kupní cena nezahr	567996	Dnes	Cihlová	Velmi dobrý
21	a1ddb463-	2017-04-28T11:3	M&M reality	e025c926-b017	2017-04-28T11	Prodej stavebního pozemku	Frenštát pod Radhoštěm,	1 130 000 Kč	1 130 000 Kč za nemovitost,	NA	NA	577141	Dnes	NA	NA
22	a1ddb463-	2017-04-28T11:3	M&M reality	af22b9d5-0257	2017-04-28T11	Prodej bytu 3+1 62 m²	Jirkovská, Chomutov	449 000 Kč	449 000 Kč za nemovitost, v	NA	S provizí RK a kon	556759	Dnes	Panelová	Velmi dobrý
23	a1ddb463-	2017-04-28T11:3	M&M reality	ba2b6f70-15ee	2017-04-28T11	Prodej stavebního pozemku	Choustníkovo Hradiště, o	950 832 Kč	950 832 Kč za nemovitost, v	NA	Cena je včetně da	577202	Dnes	NA	NA
24	a1ddb463-	2017-04-28T11:3	M&M reality	ed6374a2-4c6a	2017-04-28T11	Prodej chaty 20 m², pozem	Koryta, okres Plzeň-sever	105 000 Kč	105 000 Kč za nemovitost, v	NA	NA	575794	Dnes	Smíšená	Dobrý
25	a1ddb463-	2017-04-28T11:3	M&M reality	1aedfd2e-6779	2017-04-28T11	Prodej bytu 1+1 35 m²	Ke Statku, Dolní Lutyně	330 000 Kč	330 000 Kč za nemovitost, v	NA	cena je včetně prá	577201	Dnes	Smíšená	Velmi dobrý
26	a1ddb463-	2017-04-28T11:3	M&M reality	e136c59f-4913	2017-04-28T11	Prodej rodinného domu 16	Svojkovice, okres Rokyca	Info o ceně u RK	NA	NA	info na cenu v RK	577142	Dnes	Cihlová	Velmi dobrý
27	a1ddb463-	2017-04-28T11:3	M&M reality	615196ce-5464	2017-04-28T11	Prodej rodinného domu 14	Nádražní, Mariánské Rad	3 990 000 Kč	3 990 000 Kč za nemovitost,	NA	Cena včetně provi	577315	Dnes	Cihlová	Novostavba
28	a1ddb463-	2017-04-28T11:3	M&M reality	5431232c-6ef1	2017-04-28T11	Prodej rodinného domu 20	Otokara Březiny, Žatec	5 850 000 Kč	5 850 000 Kč za nemovitost,	NA	NA	529914	Dnes	Cihlová	Velmi dobrý

Obr. 7.4: Ukázka výstupního textového souboru pro inzeráty realitní kanceláře.



Obr. 7.5: Zobrazení polohy inzerátů největší realitní kanceláře na serveru Sreality (M&M reality).

Jako protiklad je uvedena i pobočka realitní kanceláře, která se orientuje jen na jediný region. Název této realitní kanceláře je REMAX High Way Kolín. Ukázka spuštění aplikace je na výpisu 7.3. Shrnutí použitých parametrů příkazu se nachází v tabulce 7.4. Mapa s inzeráty této pobočky je zobrazena na obrázku 7.6. Tato pobočka se nachází v Kolíně, tudíž by měla nabízet inzeráty z okolí Kolína. Podle mapy je vidět, že toto pobočka realitní kanceláře splňuje, dokonce nabízí i pár inzerátů z jiného regionu.

```
$ python3 diplomka.py --output . --singleRealEstateAgencyByName="RE/MAX
    High Way Kolín"
```

Výpis 7.3: Spuštění programu pro sbírání informací o RE/MAX High Way Kolín.

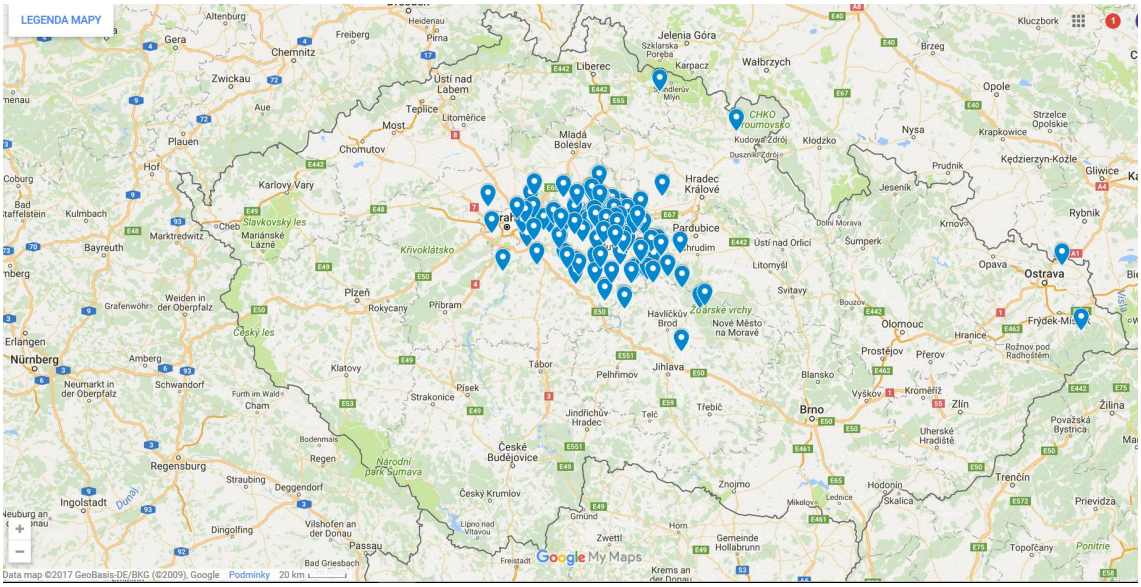
--output	.
--singleRealEstateAgencyByName	RE/MAX High Way Kolín

Tab. 7.4: Parametry pro sbírání informací o realitní kanceláři RE/MAX High Way Kolín a inzerátech, které nabízí.

Celková velikost výstupní složky bez parametru *--photo* je přibližně 6.5 GB. Běh programu bez parametru zpracování fotek trval přibližně 7 dní. Při spuštění programu s parametrem pro zpracování fotek bylo spočítáno, že program sesbíral

agencyID	fed9e356-8d65-406a-bf01-b2ff7d300f20
agencyDatetime	2017-04-17T16:55:48
agencyName	M&M reality
realityID	1ef9cca9-a6fc-4f54-9840-c4ea7350e282
realityDatetime	2017-04-28T11:31:52
realityName	Prodej bytu 3+1 70 m²
realityAddress	Záveská, Praha 10 - Hostivař
realityPrice	3 300 000 Kč
Celková cena:	3 300 000 Kč za nemovitost, včetně provize
Hypotéka:	NA
Další položky:	NA
realityBrokerName	Zákaznická linka
realityBrokerAddress	Krakovská 583/9, 11000 Praha - Nové Město
realityBrokerMobile	739 545 999
realityBrokerEmail	info@mmreality.cz
realityBrokerWeb	http://www.mmreality.cz
realityText	Textový popis inzerátu
Ulice	Záveská
Město	Praha
Okres	Hlavní město Praha
Kraj	Hlavní město Praha
Stát	Česko
SouřadniceY	50.0537358319
SouřadniceX	14.5215560754

Tab. 7.5: Ukázka výstupního souboru pro sběr informací o inzerátech.



Obr. 7.6: Zobrazení polohy inzerátů regionální realitní kanceláře (RE/MAX High Way Kolín).

přibližně 1 281 000 fotek inzerátů. Celková velikost souboru, obsahující fotky byla přibližně 24.5 GB. Kvůli vysokému počtu a velikosti fotek jsou stahovány pouze náhledy fotek, nikoli celé fotky. Průměrně vychází, že každý inzerát má okolo 13 fotek. Parametry jsou shrnuty v tabulce 7.6.

Skrip *search.py* dokáže vyhledávat nemovitosti dle zadaných požadavků uživatele. Díky tomuto skriptu je možné vyhledávat uložené inzeráty realitních kanceláří a pracovat s nimi. Díky vyhledávání podle lokality a klíčových slov, se mohou vytvářet zajímavé statistiky. Například, kolik nemovitostí v nějakém městě je připojeno na plyn atd. Na obrázku 7.7 je ukázka zobrazení všech nemovitostí na prodej, které se nachází 10 km od Brna a jejich cena se pohybuje od 2 do 3 miliónů korun. Příkaz pro spuštění programu je zobrazen na výpisu 7.4. Shrnutí použitých parametrů příkazu, se nachází v tabulce 7.7.

```
$ python3 search.py --input=slozka --city="Brno" --distance=10 --minPrice
    =2000000 --maxPrice=3000000 --type="Prodej" --output="vystup.txt"
```

Výpis 7.4: Spuštění programu s parametrem pro klíčové slovo.

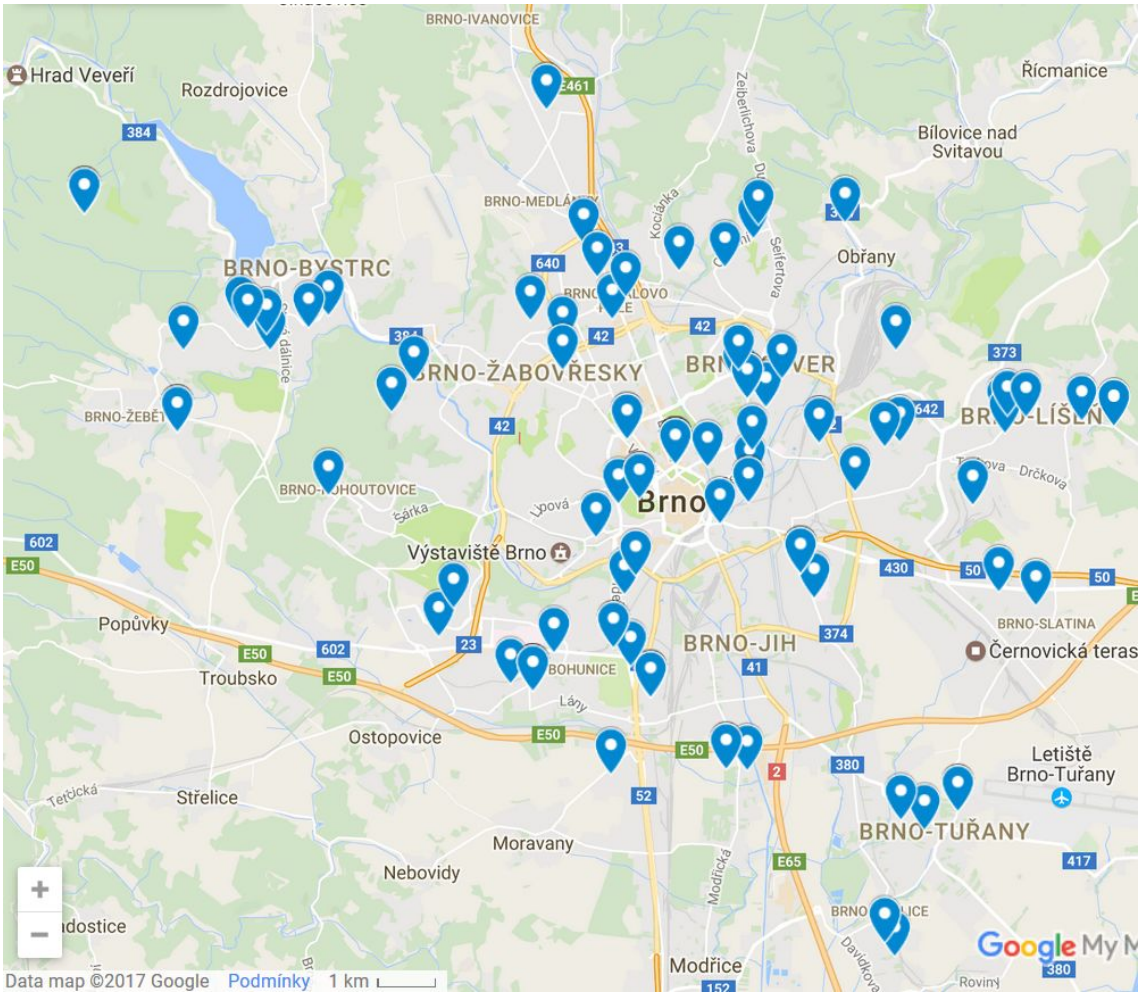
Nyní jsou z předchozích výsledků vybrány pouze ty, které obsahují sklep. Toho je docíleno tak, že je použit parametr *--keyword*. Na obrázku 7.8 jsou zobrazeny všechny nemovitosti z předchozího příkladu. Nemovitosti bez sklepa jsou označeny modrou barvou. Nemovitosti obsahující sklep jsou označeny oranžovou barvou. Spuš-

Průměrný počet realitních kanceláří	2600
Průměrná velikost souboru realEstate-Agency	650 kB
Průměrná doba sesbírání informací o RK	4 hodiny a 30 minut
Nejčastější poloha realitních kanceláří	Praha
Největší realitní kancelář	M&M reality
Průměrná velikost souboru s inzeraty	140 MB
Průměrný počet inzerátů	95 000
Úspěšnost nalezení lokality nemovitostí	99.5 %
Velikost složky bez parametru *--photo*	6.5 GB
Velikost složky s parametrem *--photo*	24.5 GB
Průměrně celkový počet fotek	1 281 000
Průměrný počet fotek na 1 inzerát	13 fotek
Celková doba běhu programu bez parametru *--photo*	7 dní
Celková doba běhu programu s parametrem *--photo*	10 dní

Tab. 7.6: Technické parametry výstupního souboru output.

--input	slozka
--city	Brno
--distance	10
--minPrice	2000000
--maxPrice	3000000
--type	Prodej
--output	vystup.txt

Tab. 7.7: Parametry pro vyhledání inzerátů na prodej, 10 km od Brna.



Obr. 7.7: Ukázka zobrazení inzerátů na prodej 10 km od Brna. Spuštěno s parametry 7.4.

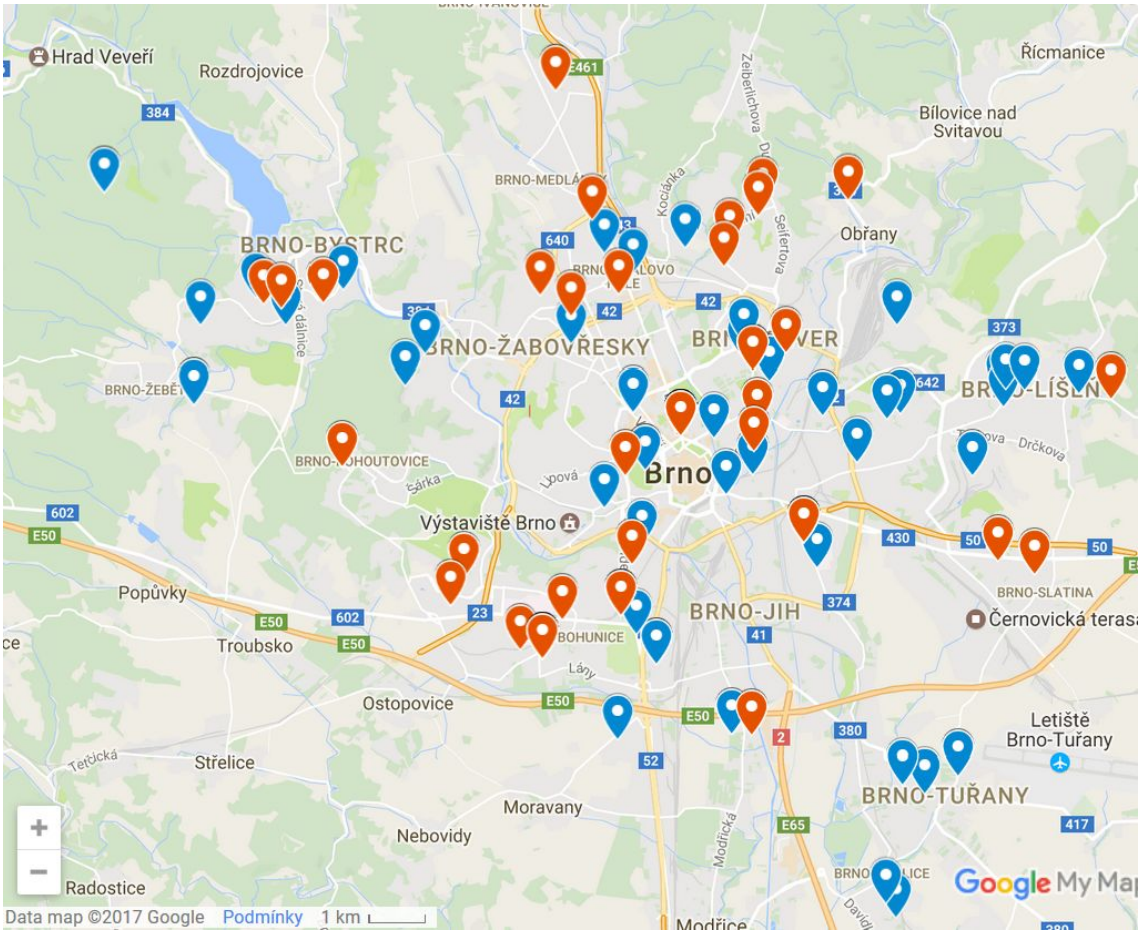
tění programu je zobrazeno ve výpisu 7.5. Shrnutí použitých parametrů příkazu se nachází v tabulce 7.8.

--input	slozka
--city	Brno
--distance	10
--minPrice	2000000
--maxPrice	3000000
--type	Prodej
--keyword	Sklep
--output	vystup.txt

Tab. 7.8: Parametry pro vyhledání inzerátů na prodej, 10 km od Brna a se sklepem.

```
$ python3 search.py --input=slozka --city="Brno" --distance=10 --minPrice
    =2000000 --maxPrice=3000000 --type="Prodej" --keyword="Sklep" --
    output="vystup.txt"
```

Výpis 7.5: Spuštění programu s parametrem pro klíčové slovo.



Obr. 7.8: Zobrazení inzerátů na prodej 10 km od Brna, obsahující sklep. Spuštěno s parametry 7.5. Modrá barva označuje předchozí inzeráty bez sklepa. Oranžová barva označuje předchozí inzeráty se sklepem.

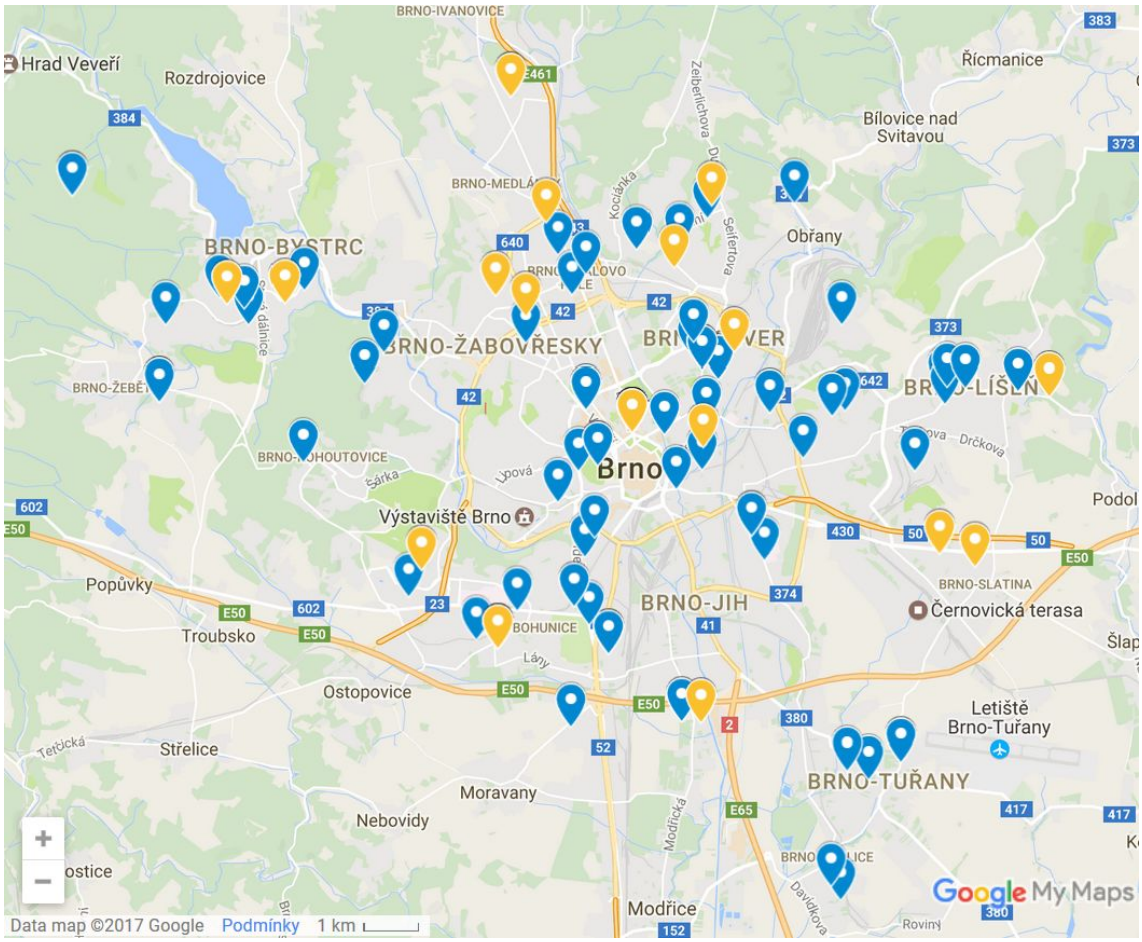
Nakonec je vyzkoušeno i reverzní klíčové slovo. Z předchozích výsledků jsou vybrány ty, které obsahují sklep a neobsahují výtah. Toho je docíleno tak, že je použit parametr *--keyword*, ale pro reverzní vyhledávání. Na obrázku 7.9 jsou zobrazeny všechny nemovitosti z prvního příkladu. Tyto inzeráty jsou označeny modrou barvou. Nemovitosti obsahující sklep a zároveň neobsahující výtah jsou označeny žlutou barvou. Spuštění programu je zobrazeno ve výpisu 7.6. Shrnutí použitých parametrů příkazu se nachází v tabulce 7.9.

```
$ python3 search.py --input=spusteni --city="Brno" --distance=10 --
    minPrice=2000000 --maxPrice=3000000 --type="Prodej" --keyword="Sklep"
     --keyword="-Výtah" --output="vystup.txt"
```

Výpis 7.6: Spuštění programu s parametrem pro reverzní klíčové slovo.

--input	slozka
--city	Brno
--distance	10
--minPrice	2000000
--maxPrice	3000000
--type	Prodej
--keyword	Sklep
--keyword	-Výtah
--output	vystup.txt

Tab. 7.9: Parametry pro vyhledání inzerátů na prodej, 10 km od Brna, se sklepem a bez výtahu.



Obr. 7.9: Zobrazení inzerátů na prodej 10 km od Brna, bez výtahu a obsahující sklep. Spuštěno s parametry 7.6. Modrá barva označuje inzeráty z prvního případu, které nesplňují podmínky pro klíčová slova. Oranžová barva označuje předchozí inzeráty se sklepem a bez výtahu.

8 ZÁVĚR

Tato práce byla vytvořena jako pomocný nástroj pro odhad ceny nemovitostí porovnávací metodou. Diplomová práce měla za cíl vytvořit aplikaci v jazyce Python, která bude prohledávat obsah webových stránek zvoleného realitního portálu v ČR. Dalším cílem této práce bylo vytvořit program pro vyhledávání jednotlivých nemovitostí podle zadaných kriterií, za účelem jejich srovnávání pro odhad ceny nemovitostí.

Úvod diplomové práce byl zaměřen na techniku dolování dat z internetu. Dále proběhlo seznámení s nástroji, jenž se využívají pro získávání dat z internetu. Před implementací aplikace byly analyzovány a specifikovány podmínky aplikace a vytvořen její návrh. Pro získávání HTML dat z dynamických stránek byl využit webdriver Selenium a knihovna BeautifulSoup. Druhá polovina práce byla zaměřena na popis implementace obou aplikací. Je zde popsáno, na jakém principu programy fungují a s jakými parametry se spouští. Na závěr jsou úspěšně otestovány obě aplikace. Výsledky testování a zhodnocení práce je popsáno v kapitole 7.

První aplikace sbírá jak data o realitních kancelářích, tak i data o jejich inzerátech. V rámci diplomové práce se podařilo posbírat údaje o všech realitních kancelářích. Získání všech dat z realitních kanceláří trvalo přibližně 4 hodiny a 30 minut, pomocí PC s připojením rychlostí 50 Mb/s, sítě VUT. V programu jsou vytvořené parametry, jejichž nastavením se může zpracovávat pouze testovací vzorek dat. Spustí se aplikace například pro prvních 100 realitních kanceláří a jejich prvních 20 inzerátů, nebo pro všechny realitní kanceláře a jejich prvních 10 inzerátů. Program pro zpracování informací o inzerátech byl otestován na vzorku vstupních dat. Důvod tohoto testování byl ten, že server obsahuje přibližně sto tisíc inzerátů a prohledávání by bylo velice časově náročné. Podařilo se několikrát získat vzorek dat 40 000 inzerátů. Z tohoto vzorku dat byly zjištěné hodnoty v tabulce 7.6.

Druhá aplikace dokáže vyhledávat nemovitosti podle zadaných kritérií uživatele. Tuto aplikaci je možné spouštět za účelem odhadu ceny nemovitostí porovnávací metodou. Aplikace pro vyhledávání nemovitostí lze spouštět se spoustou parametrů jako jsou cena, město, vzdálenost nebo klíčová slova. Tato práce vytvořila databázi nemovitostí, která by mohla napomoci k vyhledávání konkrétních nemovitostí.

LITERATURA

[1] Ing. KLIKA Pavel. *Teorie oceňování nemovitostí*. Vysoké učení technické v Brně: Ústav soudního inženýrství, 2012. ISBN 978-80-214-4567-3.

[2] BERKA, Petr. *Aplikace systémů dobývání znalostí pro analýzu medicínských dat.* [online]. 2001, poslední revize 30.5.2003, [cit. 13. 11. 2016]. Dostupné z URL: <`http://euromise.vse.cz/kdd`>.

[3] FAYYAD, Usama M. *Advances in knowledge discovery and data mining*. Menlo Park: AAAI Press, 1996. ISBN 0-262-56097-6.

[4] Michal Prokeš. *Umíte využít svá data?* [online]. 2000, [cit. 6. 11. 2016]. Dostupné z URL: <`https://www.systemonline.cz/clanky/umite-vyuzit-sva-data.htm`>.

[5] HAN, Jiawei a Micheline KAMBER. *Data mining: concepts and techniques. 2nd ed.*. London: Elsevier, 2006. Morgan Kaufmann series in data management systems. ISBN 978-1-55860-901-3.

[6] USCIANO, Chuck a Bill KENNEDY. *HTML a XHTML: kompletní průvodce*. Praha: Computer Press, 2000. ISBN 80-7226-407-9.

[7] Robert Schifreen. *The Web Book: The ultimate beginner's guide to HTML, CSS, JavaScript, PHP and MySQL*. Oakworth Business Publishing Ltd, 2010. ASIN: B007LTM67I.

[8] Richardson, L. *Beautiful Soup.* [online]. naposledy upravené 7. 11. 2016, [cit. 7. 11. 2016]. Dostupné z URL: <`http://www.crummy.com/software/BeautifulSoup/`>.

[9] Richard Lawson. *Web Scraping with Python*. Birmingham: Packt Publishing, 2015. ISBN 978-1-78216-436-4.

[10] NEGRINO, Tom a Dori SMITH. *JavaScript pro World Wide Web*. Praha: Soft-Press, 2001. Praktická vizuální příručka. ISBN 80-86497-09-7.

[11] Mitchell, Ryan. *Web scraping with Python : collecting data from the modern web*. Sebastopol, CA: O'Reilly Media, 2015. ISBN 9781491910276.

[12] Selenium Project. *Selenium Documentation.* [online]. naposledy upravené 1. 11. 2016, [cit. 10. 11. 2016]. Dostupné z URL: <`http://www.seleniumhq.org/docs/`>.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

ISO	Mezinárodní organizace pro normalizaci – pro zápis data a času
KDD	Knowledge Discovery in Databases
API	Application Programming Interface
WWW	World Wide Web
HTML	HyperText Markup Language
GPS	Global Positioning System
CSV	Comma-separated values
ipv4	Internet Protocol verze 4
ipv6	Internet Protocol verze 6

SEZNAM PŘÍLOH

A OBSAH PŘILOŽENÉHO CD

1. src/ – Složka obsahující spustitelné zdrojové soubory.
2. tex/ – Složka obsahující zdrojové soubory diplomové práce v LaTeXu.
3. tex/obrazky – Složka obsahující obrázky použité v diplomové práci.
4. diplomovaPrace.pdf – Elektronická verze diplomové práce.

B ZDROJOVÉ KÓDY

Zdrojové kódy jsou dostupné na přiloženém CD. Zdrojové kódy jsou uloženy ve složce src/. U všech zdrojových souborů se v hlavičce nachází jméno autora s popisem programu. Pro nápovědu musí být program spuštěn s parametrem *-h* nebo *--help*.

1. src/diplomka.py – Program pro vyhledávání a separování informací o realitních kancelářích a inzerátech, které nabízejí.
2. – src/localFirefox.py – Program, který stáhne, rozbalí a uloží verzi prohlížeče Firefox, na které program diplomka.py pracuje.
3. – src/search.py – Program, který ze stažených informací o inzerátech nemovitostí.

C SPUŠTĚNÍ PROGRAMU

Aplikace se spouští v příkazovém řádku. Dále jsou zobrazeny ukázky s jakými parametry lze program spouštět:

1. Získání informací o všech realitních kanceláří a všech inzerátů – *python3 diplomka.py --output . .*
2. Získání informací o 21 realitních kanceláří a od každé realitní kanceláře 10 inzerátů – *python3 diplomka.py --output . --realEstateAgency 21 --realEstate 10 .*
3. Získání informací o 21 realitních kanceláří, od každé realitní kanceláře 10 inzerátů a ke každému inzerátu fotky – *python3 diplomka.py --output . --realEstateAgency 21 --realEstate 10 --photo .*
4. Vyhledání všech inzerátů realitních kanceláří na prodej, od 2 do 3 miliónu korun a 10km od Brna – *python3 search.py --input=slozka --city="Brno"--distance=10 --minPrice=2000000 --maxPrice=3000000 --type="Prodej" --output="vystup.txt" .*
5. Vyhledání všech inzerátů realitních kanceláří na prodej, od 2 do 3 miliónu korun a 10km od Brna a obsahující sklep – *python3 search.py --input=slozka --city="Brno"--distance=10 --minPrice=2000000 --maxPrice=3000000 --type="Prodej"--keyword="Sklep"--output="vystup.txt" .*
6. Vyhledání všech inzerátů realitních kanceláří na prodej, od 2 do 3 miliónu korun a 10km od Brna, obsahující sklep a neobsahující výtah – *python3 search.py --input=slozka --city="Brno"--distance=10 --minPrice=2000000 --maxPrice=3000000 --type="Prodej"--keyword="Sklep"--keyword="-Výtah" --output="vystup.txt" .*